PREDICTIVE POWER OF CRIMINAL BACKGROUND ON LOSSES

by

Krista M. Whipple

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Business Administration

Liberty University, School of Business

December 2020

Abstract

Product and data science teams for the auto insurance industry have been trying to increase pricing segmentation with validated rating variables to decrease rate subsidization. The criminal background data availability provided a new behavior variable to test against insurance-based credit scores as a potential predictive variable in the generalized linear rating model. Criminal background was analyzed using a Poisson Log Linear model and other key insurance rating variables for predicting loss costs. The study supported the inclusion of the criminal background data in combination with insurance-based credit score as the variable's addition could improve the overall fit of the predictive model. The study also acknowledged there was a statistically significant association between criminal background and insurance-based credit score, but the overall size of the effect was small and weak. The overall contribution of value criminal background variable needs to be considered with a full rating dataset to determine if other, less powerful variables could be removed from the generalized linear to reduce the overall model complexity.

Keywords: auto insurance, criminal background, segmentation, loss cost

PREDICTIVE POWER OF CRIMINAL BACKGROUND ON LOSSES

by

Krista M. Whipple

Doctoral Study Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Business Administration

Liberty University, School of Business

December 2020

_____

Dr. David Bosch, Dissertation Chair

_____

Dr. Charles W. Williams, Dissertation Committee Member

_____

Dr. Edward M. Moore, DBA Director

_____

Dr. Anita Satterlee, Dean of Online Programs, School of Business

Dedication

For my husband, Christopher, and my son Chandler.

Acknowledgments

It takes a team to navigate the dissertation process successfully. Sincere gratitude is extended to my Dissertation Chair, Dr. Bosch, for his engagement, encouragement, and enthusiastic support throughout the process. Thank you to Dr. Michael Bates and Dr. Charles W. Williams, who offered challenging feedback, which motivated further reviews to the approach and analysis. Also, appreciation to Dr. Nicole Lowes and Dr. Edward M. Moore for ensuring the overall process went smoothly.

For my insurance team, Dr. Molly Bhattacharyya and Dr. Alex Cheng, thank you for providing the dataset as well as feedback, reviews, and modeling validation. To be able to contribute to the predictive modeling approach has been rewarding, both personally and professionally.

Table of Contents

List of Tables

List of Figures

Section 1: Foundation of the Study

The property and casualty insurance industry is rapidly changing as carriers can collect vast amounts of driver information to predict how individuals will behave while driving and paying insurance premiums (Kiviat, 2019). New usage-based insurance offerings and increased access to individual behavior variables allow carriers to better segment risks and calculate accurate premiums to address expected loss costs (Bian et al., 2018). Data-driven predictive analytics enables carriers to allocate loss costs and expenses, which results in risk segmentation (Kiviat, 2019). Advanced analytics also provides carriers with more statistical tools to handle larger numbers of complex modeling variables. The methods of sorting and ranking data the carriers use are actuarially sound, and the variables used for rating need to be classified as being fair as defined by their predictive strength in the ratemaking algorithm (Bian et al., 2018).

Background of the Problem

Auto insurance's primary role is to provide financial protection to individuals by offering a financial instrument to transfer risk with insurance premiums (David, 2015). Insurance carriers need to determine the level of risk associated with each transaction, and the amount of premium charged needs to cover the loss costs, including expenses. The optimal approach to auto insurance ratemaking and design is a risk distribution problem between the insured and the insurance carrier (Bernard et al., 2015). The advancements in financial modeling allow carriers to avoid charging the same premiums for the entire portfolio, which would underprice unfavorable risks, and as an adverse effect, would overprice standard risks (David, 2015). To mitigate adverse selection, carriers seeking growth in a highly competitive industry can no longer average price large risk pools and need to search for viable portfolio segmentation opportunities.

Cost-based pricing for risk segmentation is a traditional actuarial approach to ratemaking to match the appropriate premium to cover the associated loss costs (Segovia-Vargasa et al., 2015). Guelman (2012) stated the purpose of ratemaking approaches is to estimate future loss costs as defined as the ratio of all future claims' estimated costs against the coverage provided to cover the risk expenses and the exposure. Revising rates based on the consumer's variables will affect the profitability of a business book based on the level of premiums an individual is paying based on the risk profile (Segovia-Vargas et al., 2015). Average risk pricing and restricted rating variables eventually lead to higher rates for auto insurance costs for all consumers (Weiss et al., 2010).

One of the more important auto insurance developments was the use of insurance-based credit history data for risk classification to predict losses (Golden et al., 2016). Rating on an individual's credit history has been researched due to the controversial nature of what is considered to be a biased variable (Cather, 2018; Golden et al., 2016; Krippner, 2017). Krippner (2017) stated the credit approaches were not perfect, and the regulators have argued the use of credit-based categories would not remove the unfair treatment, as some individuals did not necessarily fit the assigned credit categories. The individual departments of insurance have considered regulatory blocks on credit for premium development, which is currently the most predictive of the auto insurance rating variables on loss cost. Private passenger auto studies have shifted away from socio-demographic underwriting factors to differentiate risks from Global Positioning Systems (GPS) to track vehicle information such as mileage, traffic conditions, and individual driving patterns (Ma et al., 2018). It is critical for auto insurance carriers to validate rating variables, and GPS rating is expensive. The majority of the financial services sector,

including insurance carriers, is more likely to invest in data and not manufactured devices, which are outside of the necessary core competencies for an insurance company (Husnjak et al., 2015).

Auto insurance applications since the 1950s have included questions concerning an individual's criminal background based on concerns of fraud and increased risky behavior. Carriers find underwriters cannot successfully act on information unless the data being provided can be verified. While the questions about past misdemeanors and felonies are asked, the data only recently became available in the United States for verification. Research through the ODEN Insurance Services Inc. regulatory database showed the individual Departments of Insurance, except for sequential rating in California, is silent on criminal background data in rating. Additional research through the Casualty Actuarial Society and recent academic insurance journals do not produce studies, supporting the verification or use of criminal background data in rating. Insurance carriers are looking to increase actuarially validated behavioral predictors of loss beyond the traditional sets of underwriting variables to create segmentation in rating and increase competitive advantages (Golden et al., 2016).

Problem Statement

The general problem to be addressed is the lack of identification and verification of highly predictive variables for auto insurance pricing and rate accuracy, resulting in the need for insurance carriers to perpetuate premium subsidization. Kang and Song (2018) stated research and development teams would need to consider several dozen rating factors for insurance modeling and predicting the target response. Selecting the top contributing predictor variables in a data set allows a modeling team to construct a regression model, with high interpretability and compelling prediction accuracy (Isotupa et al., 2019). Carriers must also mine existing company data to mirror actual customer experience to support accurate driver segmentation (Zhuang et al.,

2018). The insurance-based credit score variable is one of the most predictive variables for loss costs, and each department of insurance highly regulates the variable due to the correlation with sensitive classifications, including race and income (Morris et al., 2017). While underwriting variables outside of personal driving history are actuarially proven to be correlated to loss costs, most state departments of insurance would prefer to limit financial history rating variables in auto insurance (Morris et al., 2017). The specific problem to be addressed is the failure of auto insurance carriers to use criminal background data resulting in rate subsidization within the auto insurance industry, causing carriers to charge higher premiums for drivers with clean backgrounds and lower loss costs.

Purpose Statement

The purpose of this quantitative study is to examine if an individual's criminal background is correlated to a person's future driving behavior and if it is predictive of future loss costs. Risk segmentation attempts to mitigate pricing subsidization between lower and higher risk drivers, which supports market efficiency and addresses the increase of social risk cost and the loss of equity (Duan et al., 2018). The research by Hoy (1982) showed higher-risk classes receive coverage at an actuarially uniform premium, while lower-risk classes receive less than full coverage at an actuarially uniform premium. Hoy (1982) highlighted if the proportion of lower-risk class falls below a determined breakeven threshold, the competitive equilibrium will no longer be a no-subsidy segmentation approach. Lower-risk classes would subsidize the higher-risk classes, which would not meet management and shareholder expectations in the current economic and competitive climate.

Private passenger auto premiums serve two primary purposes for a property and casualty carrier, one being the premium should be able to cover the expense and risk obligation, and the

other should price the insurance premiums equitably for specific risk classes (Cova et al., 2016). Individual criminal background information has only recently become available as third-party data available in public and private databases. Including personal criminal background as a rating or underwriting variable would potentially allow carriers to further segment higher-risk drivers for accurate pricing, which would allow carriers to balance profitability with equitable pricing in the market.

Nature of the Study

Morgan (2018) stated the most frequently discussed means of differentiating qualitative and quantitative research designs can be determined by the data produced from the research outcomes. A quantitative research approach produces numerical data supported by validated analytical analyses, and a qualitative research approach produces results based on words (McCusker & Gunaydin, 2015). The qualitative research design will also view the research approach through a wide lens to determine patterns of relationships through an unspecified set of concepts (Morgan, 2018).

The proposed correlational design was developed to test the predictive strength of defined independent rating variables against a specific dependent variable, pure premium, which is defined as being the estimated incurred losses divided by the earned car year exposures for a book of business (Frees et al., 2016). Correlational research is focused on explaining relationships between two or more variables in one or two populations (Curtis et al., 2016). Quantitative researchers want to establish why and how variables differ and determine how one independent variables' variance may be associated with the variances in another independent variable (Curtis et al., 2016).

Discussion of Method

Creswell and Creswell (2018) stated research designs are the methods, which connect the steps from broader assumptions to detailed procedures for aggregating data and conducting the appropriate analysis. A quantitative approach for the proposed research design was selected based on the requirements for creating a formal path to providing replicable numerical outcomes. For quantitative research approaches, variables are isolated and defined by categories, which can frame hypotheses before the data being collected and used for testing and modeling (Brannon, 2016).

The study aims to determine the predictive strength of independent rating variables for auto insurance ratemaking purposes. A qualitative approach was not selected as being an appropriate method for the study because the independent variables require verification and cannot be manipulated (Curtis et al., 2016). Descriptive quantitative approaches are more appropriate for the social sciences, where verifiable data are not as readily available for analysis (Siedlecki, 2020). The qualitative approach is not a hypothesis testing design, so there are no independent or dependent variables, and there are only variables of interest (Siedlecki, 2020).

Discussion of Design

Casual observation offers an approach to quantitative research, which underscores design-based inference methods (Imbens & Rubin, 2015). The causal-comparative design tries to influence specific causal facts for defined subpopulations and is associated with identification strategy research designs (Samii, 2016). The research approach strives to find relationships between the dependent and independent variables after events have taken place.

The correlation design is better aligned with the proposed study, as there are specific hypotheses to test and is concentrated on variances in variable relationships. A quantitative

research design is an objective and systematic process to define variables and test relationships for potential correlations between the variables (Bloomfield & Fisher, 2019). Quantitative approaches align well with financial services and pricing research as the methodology strives to find an exact answer to a hypothesis using objective and balanced scientific methodologies (Bosco et al., 2015).

Summary of the Nature of the Study

In property and casualty insurance, quantitative research provides an opportunity to work with extensive collections of numerical data sets using statistical measurements and outcomes to classify relationships and patterns within the data. Multivariate classification ratemaking has rapidly advanced during the past ten years, which allows for different types of statistical approaches for segmenting and pricing individual risks (Miljkovic & Fernândez, 2018). Predictive modeling affords insurance carriers the critical advances of the equitable pricing of risks, a better competitive advantage, and protection from adverse selection, allowing for better-informed decision-making driven by verified data (McCusker & Gunaydin, 2015).

Research Questions

Multivariate classification ratemaking has rapidly advanced during the past ten years, which allows for different types of statistical approaches for segmenting and pricing individual risks (Miljkovic & Fernândez, 2018). Predictive modeling affords insurance carriers the critical advances of the equitable pricing of risks, a better competitive advantage, and protection from adverse selection, allowing for better-informed decision-making driven by verified data (Shi et al., 2015). Miljkovic and Fernândez (2018) stated predictive modeling is used for risk classification for individual risks. At the aggregate level, the predictive models can assist with quantifying risk segments of a portfolio.

Current auto insurance pricing approaches use generalized linear modeling with the most common approach to handling zero inflation being the frequency-severity model and the Tweedie compound Poisson model (Shi et al., 2015). With the rapid advancements in data collection and open-source statistical code, carriers are now able to collect large amounts of external and internal data to support more advanced modeling (Kafková & Křivánková, 2014).

RQ1. What is the predictive impact of an individual's criminal background on auto insurance loss costs?

RQ2. What is the relationship between the insurance-based credit score and criminal background?

RQ2.a. What is the outcome of the predictive model if the insurance-based score is removed and is replaced with the criminal background variable?

Hypotheses

A property and casualty insurance carrier was identified to provide a modeling data set with appended criminal background data. A Data Science Team will ensure the data set is sequestered and validated. A Chi-Square Test will determine the association between the predictor variables (Krzywinski & Altman, 2015). The multiple linear regression is practical for estimating the properties of predictor variables, and the estimated regression coefficients are dependent on the predictors in the model (Fang et al., 2016). Future behavior is challenging to predict, and individual risk models are developed and calculated by historical loss events (Krzywinski & Altman, 2015). The most predictive variables in the current Generalized Linear Model will be used for the study's variable testing. The coefficient of determination, R squared, will determine the degree of interrelation and dependence between the variables (Krzywinski & Altman, 2015).

Data Set

A random sample of approximately 448,755 auto insurance records from a cleansed, normalized data set of non-standard auto risks from 2014 to 2017 will be appended with third-party driver-level criminal background data. A data test with a random sample of drivers over five years was run with vendor data to determine the hit rate and rating lift potential. The countrywide driver criminal background hit rate was approximately eight percent with a loss ratio, the percentage of total claims paid with the total earned premiums, 10 – 20 points higher than the average loss ratio for the business. The appended data set was researched and tested by independent actuarial groups for validity and reliability. The normalization of the data set allows for control and accuracy throughout the analysis (Zhu et al., 2017). For modeling purposes, a 75/25 split for a training data/test approach will validate the hypothesis testing for each of the research questions. For insurance ratemaking, 2,890 incurred claims are required for full credibility (Casualty Actuarial Society, 1990). Pure premium is the incurred losses and loss adjustment expenses divided by earned car years or frequency multiplied by severity, which is the dependent variable for each model to determine the expected value of the outcome for insurance loss costs (Werner & Modlin, 2016). The predictor variables are the current rated policy variables, which a carrier includes in the filed class plans along with the appended driver background variable (Kafková & Křivánková, 2014). Each model's predictor variables will target loss frequency and average loss amounts for continuous and categorical variables.

$H1_o$ = There is no additional incremental predictive ability of the current pure premium (DV) model with the addition of the criminal background data.

$H1_1$ = There is additional incremental predictive ability of the current pure premium (DV) model with the addition of the criminal background data.

The ability to segment risk elements supports insurance carriers with appropriate risk level pricing and mitigates overcharging risks, which should be paying lower premiums for less risky driving behavior (Pechon et al., 2019). Verification of insurance rating variables with third party data is also costly and adds to the expense ratio. If verified data are not significant, the new variable will not be supported by the business case and rejected as credible support for ratemaking (Porrini, 2015).

$H2_o$ = There is no statistically significant correlation between an individual's criminal background and insurance-based score.

$H2_1$ = There is a statistically significant correlation between an individual's criminal background and insurance-based score.

The state departments of insurance require clear analytical support of predictive variables used in ratemaking (Ranganathan et al., 2017). The criminal background variable needs to be reviewed for predictive strength, along with the current rating variables, to determine the level of significance. The approach should also determine if there is any interaction between the most predictive variable, credit score, and the individual's criminal background and if the independent variables have multicollinearity affecting the accuracy of the regression model accurate (Pechon et al., 2019).

$H2a_o$ = There is no increase in the incremental predictive ability of pure the premium model with the addition of the criminal background data and the removal of the insurance-based score variable.

$H2a_1$ = There is an increase in the incremental predictive ability of the pure premium model with the addition of the criminal background data and the removal of the insurance-based score variable.

Property and casualty carriers use insurance-based scores in underwriting and rating, and the practice is controversial, with many consumer groups disagreeing with the use of credit for premium development (Morris et al., 2017). Finding a more predictive variable for pricing segmentation would provide carriers more flexibility in ratemaking and also provide alternatives to large markets, such as California ranked as the largest auto insurance market as reported by the National Association of Insurance Commissioners, which have banned the use of credit scores for premium development (National Association of Insurance Commissioners, 2019). Some departments of insurance, and consumers, may find validated criminal background may not correlate with questionable classifications (Morris et al., 2017).

Theoretical Framework

In the current fast-paced and growing data environment, auto insurance ratemaking based on business analytics allows carriers to develop premiums using non-traditional rating and underwriting information (Frees et al., 2014). The auto industry defines rating variables such as insurance-based scores, prior insurance bodily limits, and homeownership as non-traditional as they do not conform to the traditional descriptive characteristics for the vehicle or the individual (Pechon et al., 2019). Carriers who fail to further segment risks with strong predictive variables can be subject to adverse selection in the marketplace leading to the unintended retention of lower premium policyholders with higher associated loss costs.

State Regulations

Insurance carriers are required to justify their premium development approaches and profitability targets for business lines using loss and expense controls (Kimball & Boyce, 1958). Auto insurance ratemaking is regulated in approximately half the states, with the other half being classified as more of an open and competitive environment (Weiss & Choi, 2008). Challenges

with the execution of standardized ratemaking led to a constant cycle of new product development, which required regulatory overview at the state level (Kimball & Boyce, 1958). Regulation of insurance rates at the state level became a more advanced solution, and Congress enacted the Mccarran-Ferguson Act declaring continued regulation and taxations by the states for the business of insurance was of the public interest (Weiss & Choi, 2008).

Application Variables

For property and casualty insurance carriers to adequately price the risks, ratemaking relies on accurate estimation of future loss costs associated with the coverage provided (Garrido et al., 2016). The modern modeling approaches for auto insurance pricing methodology are developed using generalized linear models because the means of the frequency and severity processes are expressed through linear combinations of rating variables found in the traditional insurance application for an insurance contract (Quijano-Xacur & Garrido, 2015). Traditional application rating variables such as household drivers, vehicles, traffic violations, and accidents are used by all carriers and are the baseline point for rating. Those variables are the starting point, and innovative carriers seek new, validated rating variables to differentiate the products in the marketplace.

Actuarial Ratemaking

Credibility theory is one of the cornerstones in actuarial science and is one of the highest accuracy theories based on the Bayesian interpretation of probability (Xie et al., 2018). Linear models have limited application in actuarial science stemming from the fat the insurance data are right-skewed, or discrete from the claims frequency models, therefore the linear models need to be broadened to Generalized Linear Modeling (Pechon et al., 2018). Using Generalized Linear Modeling, premium development is no longer reliant on normal distribution assumptions and can

be more flexible and include distributions using exponential modeling such as Poisson and Gamma distributions (Xie et al., 2018). Advancements in statistical modeling, supported by increased computing power, allow for variable testing using multiple linear regression and Generalized Linear Modeling. For state regulatory filings, both approaches to linear modeling are accepted by state actuarial departments.

Verified Third-Party Data

Auto insurance producers seek to place risks with the right carrier at the best available premium in the market. Using inaccurate data for developing rating models or pricing can lead to revenue loss, process inefficiencies, and potentially the inability to comply with insurance statutes and regulations (Gao et al., 2016). Unverified data provides producers and insureds avenues to create pricing gaps in ratemaking and leads to average pricing across pools of risks instead of providing the right rate for each risk. Unless rating variables are verified, the carrier cannot use those variables for meaningful rate segmentation, and it is unlikely the respective departments of insurance would approve filings without the appropriate actuarial exhibits.

Premium Subsidization

The analysis of premium subsidies emphasizes increased insurance demand compounds moral hazard in the market, and the effects have been mostly silent in supporting literature (Jaspersen & Richter, 2015). The markets studied have been open insurance markets, focusing on the demand effects of subsidies and the associated assumption models (Jaspersen & Richter, 2015). Many regulators and consumers view auto insurance pricing as a win-lose relationship, and the one entity which gains from the pricing advantage causes the other party to lose (Hinterhuber & Liouzu, 2017). In reality, subsidies alter the structure of the insurance contract as they reduce the premiums and increase the wealth of the insured (Hinterhuber & Liouzu, 2017).

Outside of the insurance structure's specific design, premium subsidies need to be financed, and the two logical choices in auto insurance are the larger risk population or the specific group driving the loss costs. The current market structure does not allow for carrier subsidization due to the tight margins and shareholder return on equity requirements.

Modeling

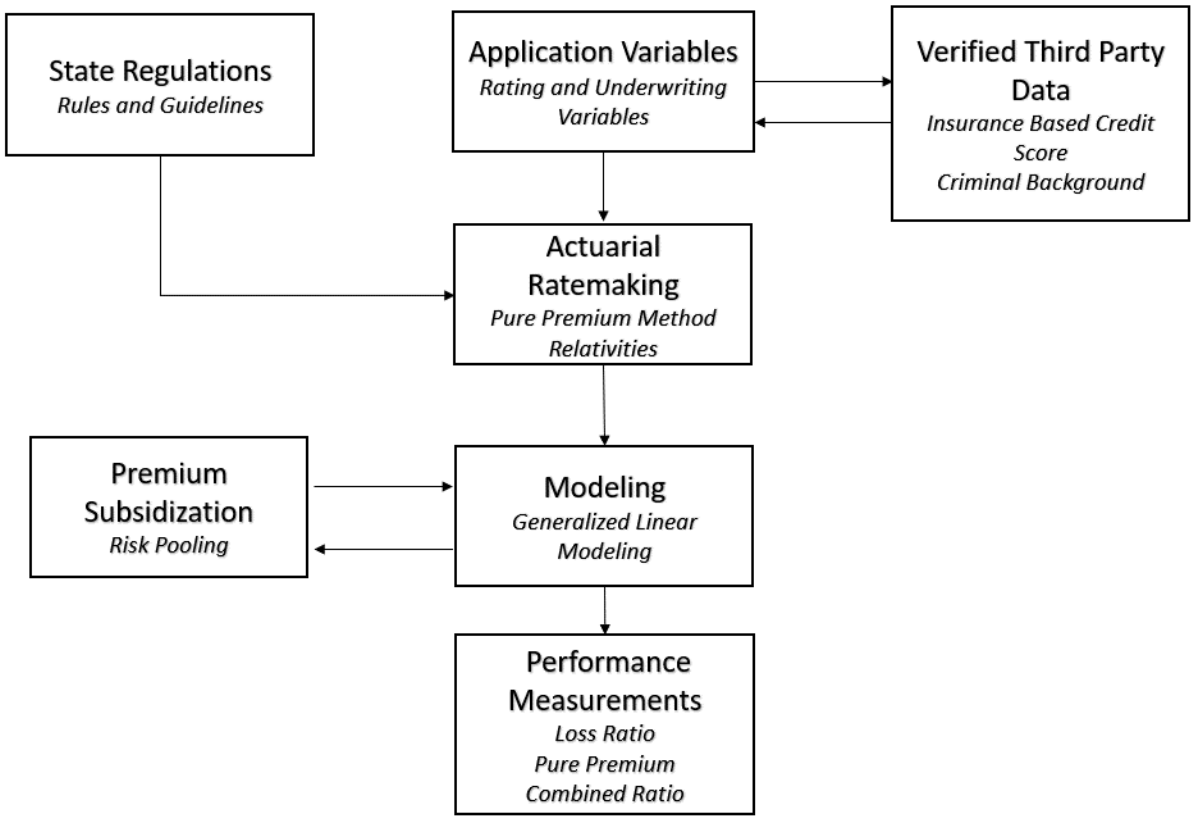
The Generalized Linear Model accounts for dependence in a straightforward approach, which is accepted by the individual departments of insurance and is also easy to implement and explain to executive management teams (Garrido et al., 2016). Additionally, the total loss cost can also be modeled directly with the Tweedie distribution, which uses the aggregate claims as a compound Poisson-Game sum and assumption of independence between claim counts and claim size (Quijano-Xacur & Garrido, 2015). The predictive variables can be evaluated with multiple linear regression for strength, and then integrated into Generalized Linear Models and Tweedie distributions for overall model contributions.

Performance Measurements

Premium, losses, and expenses are the variables used to determine the property and casualty insurance carrier's overall profitability. The combined ratio measures the underwriting profit of a business line for a carrier and aggregates the loss ratio and the expense ratio (Graham & Xie, 2007). A combined ratio under 100 is considered to be an underwriting profit. The loss adjustment expenses divided by net earned premiums defines the loss ratio, and underwriting expenses divided by net written premiums defines the expense ratio.

Figure 1

Relationships Between Theories and Variables



Discussion of Relationships Between Theories and Variables

The study's theoretical framework will test the relationship between criminal background and insurance-based credit scores on pure premium or loss costs. Criminal background has not been used in auto insurance premium development and has only recently been verified by a third-party vendor. The independent variables, criminal background, and insurance-based credit score will be modeled with the dependent variable pure premium to determine if a predictive relationship exists. Modifying variables would include driver violations and accidents.

Summary of the Conceptual Framework

Prior literature supports validated third-party variables for auto insurance ratemaking (Lemaire et al., 2015). Current searches of auto insurance carrier filings with the individual

departments of insurance and current literature are silent on verifying criminal background as a predictor of future loss costs in ratemaking. The modeling and validation approaches are within the acceptable actuarial science methodology guidelines and can be supported with the necessary actuarial exhibits for rate filings fulfilling regulatory guidelines. Criminal background for additional segmentation in ratemaking further mitigates rate subsidization from the pool of auto insurance risks and rates the individual risk with better accuracy.

Definition of Terms

There are industry terms and definitions, which are important for the research study concerning auto insurance ratemaking. In auto insurance, the financial performance relationship is expressed in pricing and various profit measurements (Shim, 2017). The dependent variable, pure premium, the key independent variables, insurance-based credit score, and criminal background, along with central insurance financial performance variables, have been reviewed in additional detail. Any further industry acronyms will be fully defined throughout the study, if necessary.

Combined Ratio, Performance Measure. The combined ratio adds the percentages of the loss ratio and expense ratio to determine the overall business profitability (Graham & Xie, 2007). Expenses include loss adjustment expenses and underwriting costs. A carrier with a combined ratio of over 100 percent is not profitable and requires either loss or expense reductions to maintain a target ratio acceptable to management and shareholders.

Criminal Background, Independent Variable. Criminal background is defined as the most recent eight-to-ten years of an individual's misdemeanor and felony convictions records. An individual's criminal records do not include convictions for driving violations or accidents, and

there is no overlap with driving backing and an individual's criminal history. Only convictions are considered as part of an individual's criminal background.

Insurance-Based Credit Score, Independent Variable. Fair and Isaac developed credit scoring in the early 1960s with the base algorithm supporting the estimation of risks as related to credit products using a consumer's personal information such as annual income, occupation, and overall financial responsibility (Livieris et al., 2018). Credit scoring became one of the most noteworthy and successful operations research methods and was adapted by the insurance industry to identify those consumers who exhibit more risky behavior (Livieris et al., 2018). An insured-based credit score is calculated from public and private consumer variables, representing the probability of an insured filing a loss or paying their premiums on time (Kiviat, 2019).

There is a strong correlation between insurance-based credit scores and incurred losses, verified by multiple studies with no study reporting a lack of a strong statistical relationship being published in the peer-reviewed literature (Ahlgrim & Jones, 2014). Of the studies conducted, an insurance-based credit score is in the top three most predictive variables and is most often the most reliable for liability coverages (Brockett & Golden, 2007). The use of insurance-based credit scores is controversial, and departments of insurance would prefer to eliminate the rating variable due to the challenge of determining why there is a correlation between scores and incurred losses.

Modern practices of using insurance-based credit scores for ratemaking have also come to social science researchers' attention, concerned with the topic of risk management in consumer finance (Gennaioli et al., 2015). Gennaioli et al. (2015) argued risk management measurement systems are unstable and require constant maintenance with new data to maintain their predictive power. The current systems can refresh data daily, and social scientists and

insurance regulators do not acknowledge if individual risky financial decision-making translates to risky driving behavior even with the proper actuarial and statistical support.

Loss Ratio, Performance Measure. A carrier's auto insurance loss ratio is the relationship between incurred losses and earned premiums (Graham & Xie, 2007). Loss ratios are calculated as a percentage of the incurred losses divided by earned premiums. Incurred losses are paid claims, including loss reserves. Loss reserves are liabilities for known losses, which have not been paid by the carrier. Earned premiums are the portion of written premium earned during the policy period.

Pure Premium, Dependent Variable. Pure premium estimates the incurred losses, including loss adjustment expenses divided by the earned car year exposures (Graham & Xie, 2007). The pure premium can also be expressed as frequency multiplied by severity. The statistical approaches will allow a researcher to model frequency and severity with different statistical models or with pure premium in one model.

Assumptions, Limitations, Delimitations

When insurance carriers face the challenge of pricing insurance risks, there are usually unrealistic assumptions that various types of claims events are independent (Bermúdez et al., 2018). With advancements in data collection, computing power, and model sophistication, researches have verified the positive correlation between claims types and introduced multivariate regression models to ease the independence assumption between claims counts from a policy (Bermúdez et al., 2018). The analytical and data developments also allow carriers access to more information to augment data sets and expand forecasting and modeling capabilities for more flexibility in ratemaking approaches.

Assumptions

Private passenger auto carriers use multiple linear regression models to accomplish accepted a priori ratemaking approaches for insurance modeling (Baumgartner et al., 2015). Morata (2009) stated when assuming independence between claims events, the rates could be calculated by adding the premiums for each line coverage, dependent on the rating factors selected. A path to working through the concern of heterogeneity for a priori ratemaking includes segmenting risks into homogeneous rate classes, so the insureds belonging to a specific class are paying similar rates (Antonio & Beirlant, 2007). For auto insurance, grouping the risks into homogeneous classes can be accomplished using classifications variables such as insured-based credit score or criminal background because the values can be determined at the start of the ratemaking process. Different types of multiple regression models will be investigated, so in some instances, the independence assumption will be relaxed for additional flexibility in testing approaches.

An assumption will be made to account for long-tailed severity distributions that the marginal distribution follows a generalized linear model (Frees et al., 2016). A generalized linear model allows for selecting the dependent variable distribution and permits the inclusion of explanatory variables in the normal modeling process (Antonio & Beirlant, 2007). For auto insurance ratemaking, the generalized linear model approach can be used for continuous and discrete results (Frees et al., 2016).

Limitations

The study's primary limitation is the recent introduction of the verified criminal background data and the shorter development period for losses being analyzed with the dependent variable, pure premium. The pure premium is the average incurred loss by exposure

unit and equals the product of the frequency of claims events per earned car year and claims severity. An earned car year is the time in units, exposed to loss during the policy period. In property and casualty insurance, claims sometimes require several years to be settled, and seven years of data would have been preferred for modeling purposes (Denuit & Trufin, 2018). Increasing the population sample to the maximum number of policy records for the states with available reduces the standard error and increases the credibility for the proposed research.

The study's data extract is from a private passenger auto carrier and includes data from the exposure period January 2017 through June 2019 evaluated in February 2020. Modeling data are historical, developed loss data, and excludes any expense or catastrophic information. The proprietary insurance-based credit scores are provided by a third-party vendor, which are grouped, eliminating any association with personal identification information. A potential observed limitation may be removing personal identification information, which is a normal process for insurance carriers and is not used in the modeling or rating process (Denuit & Trufin, 2018). The removal of personal identification information protects an individual's sensitive information, and it also helps mitigate perceived biases where rating information may support ethnic or gender profiling.

Delimitations

The dataset was limited to the most influential predictive variables used in the generalized linear model used for auto insurance ratemaking. While all of the variables in the current generalized linear model have positive values, the study's modeling iterations can be better managed with a streamlined dataset without losing model integrity. The variables with the most significant contributions within the model will help determine if the new variable being introduced contributes to the model's overall fit for rating accuracy (Ajiferuke & Famoye, 2015).

Because California is a highly-regulated premium-subsidization model, it will not be included in the dataset. The California Department of Insurance prescribes the rating and underwriting variables allowed by auto carriers, including the sequential rating model for developing premiums. The approval of Proposition 103 also mandated the prior rate approval process increasing review times and decreasing rate segmentation by allowing consumer intervention in the insurance pricing process. The objective was designed to reduce rate differences between drivers. Instead, the proposition resulted in the higher-priced nonstandard market becoming more populated and increased the need for rate subsidization by the standard markets (Ippolito, 1979).

The removal of personal identification information through de-identification is necessary to ensure information is not inadvertently linked back to the consumers who own the data (Garfunkel, 2015). When variables contain identifying information such as policy numbers, names, or geolocation information, there may be a conflict between the intended goals of the underlying data being used and privacy protection (Garfunkel, 2015). No policy numbers, names, or zip codes were included in the dataset. Insurance-based credit scores were coded to groups, and the individual scores were not included in the dataset for analysis purposes.

Significance of the Study

Cather (2018) stated there is a recurring theme in scholarly research concerning property and casualty insurance, where adverse selection from average pricing risk pools can create a pricing event, which left unchanged, may cause a negative financial impact on a carrier. Adverse selection occurs when an insurance carrier's earned premiums cannot cover the incurred losses due to underpricing high-risk consumers by failing to identify the predictive variables contributing to the loss costs. Adverse selection affects carriers relying on classifications to

segment risks into pricing categories with similar characteristics (Cohen & Siegelman, 2010). The insurance-based credit score was the last significant rating variable defined in the 1990s, which is a derivation from traditional credit scores designed for use in the insurance industry, and are highly controversial. Consumer and legal groups continue to pressure the state departments of insurance to have insurance-based credit scores reduced or removed entirely.

Telematics is the technology, which allows carriers to collect navigation, safety, and driving behavior information through electronic devices (Ayuso et al., 2019). Full telematics systems are expensive and challenging to integrate into the auto insurance operations model, even with smartphone technology, including accelerometers. Fundamental changes in-vehicle data sharing, sales, operations, and pricing in the automobile and insurance industries would need to occur before the widespread adoption of telematics programs (Ayuso et al., 2019). The introduction of the criminal background data may provide an additional predictive variable to property and casualty carriers who do not have the resources to invest in full telematics programs and are looking for acceptable options outside of an insurance-based credit score.

Reduction of Gaps

The study intends to research and potentially validate personal criminal background as a predictive behavior variable, which may help segment risks for accurate premium development at the individual risk level. Risk classification for property and casualty insurance is faced with internal and external challenges when examining segmentation variables due to conflicts in efficiency, equity, and social objectivity (Aseervatham et al., 2016). Rating variables need to be observable and in data-driven carriers, verifiable, to adequately price and structure the risk (Cohen & Siegelman, 2010). Carriers strive to reduce data irregularities by collecting, validating, and classifying applications to develop individual premiums, which at times cause public and

insurance department concern (Aseervatham et al., 2016). From the carrier perspective, unverified variables lead to premium leakage, overall average pricing, and decreased differentiation due to underrepresentation of the risk at the time of rating.

Insurance regulators and state legislations are inclined to propose and implement restrictions for segmentation variables to prohibit those rating, and underwriting variables deemed discriminatory (Thiery & Schoubroeck, 2006). The task of providing support and the actuarial verification of the new variables which are not biased or discriminatory falls with the carriers and the associated statistical exhibits. The decision to introduce additional risk classification through rate segmentation requires a well-funded and statistically significant justification for department rate filings (Cummins & Tennyson, 1992). The path for determining a differentiation approach should bridge the conflict between an individualistic moral rights approach and an insurance group approach to equity (Thiery & Schoubroeck, 2006).

Implications for Biblical Integration

Researchers have been developing studies to define ethics and morals displayed in certain types of individual consumer behavior (Cova et al., 2016). Some regulators and legislators would prefer insurance carriers treat individuals as a generic pool of risks and ignore specific identifying characteristics. For criminal background, a subsidization approach would effectively ignore the risky behavior displayed by a group of individuals and penalize clean individuals creating a financial imbalance. Johnson et al. (2016) noted the Christian interpretation of redistributive guidelines and social justice is not always consistent due to the blurred definition of what is considered to be justice.

The implications are while some people may argue social justice is based on human rights, there is a difference between selecting to engage in risky behavior and working through

economic injustices. Ishida et al. (2016) stated insurers had made significant devising processes, which effectively target probable risky and fraudulent claims related to actions with statistical modeling and business intelligence software. Applying situation-specific moral intensity to rating variables and underwriting criteria can be critical in pricing outcomes because it creates a bridge between appropriate moral conduct and personal behavior choices, supported by both regulators and consumers (Ishida et al., 2016).

Relationship to Field of Study

Cummins and Xie (2016) highlighted insurance was constructed of several concepts. The two most notable are the accurate statistical models of insurance risk pools originating from probability theory and actuarial science and the insurance firm's financial models and the pricing, derived from financial theory (Cummins & Xi, 2016). Insurance demand theory and the application of different analytical approaches are grounded in economics, and the components of modeling and financial theory create the basis of the field of finance and insurance economics (Mankaï & Belgacem, 2016). Auto insurance is explicitly a method of allocating different types of financial risk and providing services associated managing the risk (Cummins & Xie, 2016).

The original financial models integrated insurance rating variables into an economic context with many models resulting in financial market equilibrium or an attempt to mitigate an arbitrage position (Lee et al., 2005). A step toward integrating finance, insurance, and statistical theory is by using models, which incorporate realistic claims distributions with predictive variables into the financial ratios used to set rates for insurance products (Lee et al., 2005). Auto insurance is driven by analytical competition, and the choice of an appropriate business model is a critical business and financial decision (Kim & Min, 2015). A carrier's business model is an essential driver of innovation and a source of value creation for the organization and the

stakeholders (Soleymanian et al., 2019). As insurance carriers create new products and services, product research and development teams need to develop new business models to realign their systems and processes to support the influx of new data and information (Kim & Min, 2015).

Summary of the Significance of the Study

Including criminal background as a potential variable for predicting auto insurance loss costs could provide a distinctive pricing capability reducing the need for average pricing and premium subsidization by auto insurance carriers. The study will reduce the literature and research gap for predictive variables and better define the impact of misdemeanor and felony activities on individual risk assessment. In current academic literature, there has been no mention of the verification or use of criminal background behavior for rating or underwriting for auto insurance premium development. The research may also provide moral paths to help Christians better understand the social justice for pricing risks based on individual behavior and choices.

Financial modeling progressions allow insurance carriers to avoid charging an average premium level for the entire risk portfolio, which underprices unfavorable risks and overprices standard risks. The majority of carriers have access to similar data, barring usage-based information, and any new information provides segmentation and competitive advantage over other auto insurance carriers. With a countrywide hit rate of eight percent, carriers can either price accurately for those higher frequency risks or send those risks to another carrier. The remaining risks would be afforded a lower total premium to match their actual driving experience, attracting the more profitable lower-frequency risk to support profitability and product growth.

A Review of the Professional and Academic Literature

The property and casualty insurance ratemaking process is an experience rating system, which uses an insured's driving record to determine future policy premiums (Najafabadi et al., 2017). Some of the basic actuarial approaches originated from studies of claim count frequency, which was believed to frame the overall risk classification rather than claim severity or the claim's overall cost (Jeong et al., 2017). Ratemaking takes into consideration the frequency per earned exposure to determine premiums for individual risks.

For modern ratemaking, the insurance industry uses a bonus-malus contract system, which allows for specific results based on either positive or negative outcomes (Najafabadi et al., 2017). A bonus-malus rating approach for auto insurance is only based on the claims events modifying the premium development (Gómez-Déniz, 2016). The pure premium calculation has been based on modeling both the frequency and severity of loss events in a combined claims model (Jeong et al., 2017). Actuaries have a customary practice of assuming independence when using frequency and severity for premium determination using loss costs (Najafabadi et al., 2017).

As additional data elements become available for statistical modeling, carriers can observe more and new dependence patterns, which had not previously attracted enough attention due to the relevant data's non-accessibility (Hua, 2015). The increased computing power and advancement in statistical approaches have made advanced models available, which describe new dependence patterns and are easier to understand and implement for ratemaking and department of insurance filings (Jeong et al., 2017; Jeong et al., 2018). The cutting-edge statistical modeling allows for testing new variables and provides insights on potential predictors, including observed individual behaviors leading to increased loss events.

Actuarial Ratemaking

The bonus-malus approach for insurance pricing was developed while studying other auto insurance rating variables such as age and sex of the driver as well as vehicle attributes and the territory where the risk is underwritten (Lemaire et al., 2015). Lemaire et al. (2015) explained all *a priori* and *a posteriori* classification variables should be incorporated and validated for the model to ensure drivers are not penalized with higher surcharges and cumulating an *a priori* premium increase. Consumers and departments of insurance have valid concerns of double-counting similar effects due to certain variables being associated with higher claims propensity, and a bonus-malus approach is necessary to avoid excessive penalties for riskier drivers (Gómez-Déniz, 2016).

Several methodologies can predict the expected number of loss events, and traditional linear modeling can be used to identify significant risk classification variables, determine tariff classes, and develop premiums (Dionne & Vanasse, 1992). An accurate ratemaking approach allows insurance carriers to cover expected losses and expenses using modeling, which describes the claims frequency distribution. Yip and Yau (2005) stated the claims count distribution in auto insurance ratemaking is assumed to follow the Poisson and negative binomial distributions. Poisson and negative binomial models with a regression component incorporate all available loss information to accurately predict accident distributions (Dionne & Vanasse, 1992). A negative binomial distribution with applied regression elements provides a reasonable estimation of the actual loss event distribution (Yip & Yau, 2005). Under the deductible agreement in the auto insurance contract, a claim will not be opened if the amount claimed is small, creating an excess of zero claims. Eryilmaz (2016) and Gómez-Déniz (2016) highlighted the modeling of claim frequency distributions using compound models. The zero-inflated Poisson model is receiving

more interest as the approach considers discrete count data in order to allow for the occurrence of excess zeros. The method is widely used to streamline the claims estimation process and increase prediction accuracy in the industry.

Rating. Auto insurance premium development can be completed in two phases, with *a priori* rating in the first phase, where carriers use verified and observed risk classification variables to segment a risk pool into homogeneous risk classes (Tan, 2016a). While the *a priori* risk characteristics are important to all carriers, they do not complete the full risk profile and require new *a priori* variables along with *a posteriori* rating in the second phase (Tan, 2016a). Under a bonus-malus system, the framework of credibility premium is established to address the residual heterogeneity based on claims experience information, which is considered an unobservable risk only provided by the individual's loss history (Gómez-Déniz, 2016).

Credibility theory is the merging different groups of data sets to obtain an accurate overall estimate and provides actuaries with methods to develop insurance premiums for a heterogeneous book of business (Najafabadi et al., 2017). Credibility theory is a group of quantitative approaches, which allow carriers to model future premium patterns based on historical experience and is the weighted sum of the sample mean and the written premiums (Gómez-Déniz, 2008). The calculated weighted factor is the credibility factor used in the experience rating (Karmila et al., 2020). Payandeh and Amir (2010) stated several different approaches could lead to the same credibility factor expression, such as the distribution-free method and the Bayesian methods.

Historical Modeling and Underwriting. Boyer and Owadally (2015) stated the existence of profitability cycles based on linear time series analysis was proven to be present in property and casualty insurance. Property and casualty insurance uses nonfinancial models for

pricing, and the Sandmo-Leland model is based on the assumption the carrier takes a risk-averse position and suggests the price of a policy equals the expected cost of the policy plus a risk premium (Choi et al., 2002). Researchers also agree there is little evidence supporting the fact the insurance cycle follows a linear autoregressive pattern, which means any cyclicality in carrier profitability is not predictable as defined by a traditional economic framework (Boyer & Owadally, 2015). The autoregressive process is a type of statistical modeling, which provides a fundamental interpretation of insurance market performance, does not provide adequate representation, and is not the only available description for insurance cycles.

Regulatory Environment. Schwarcz (2018) highlighted property and casualty insurance carriers must comply with a complicated and restrictive state-based regulatory system, which the controlling body prohibiting what each department considers excessive and unfair rates. The current regulations evolved from a group of regulatory and market conditions, which no longer exist in most of the property and casualty insurance markets. The continuation of the traditional insurance rate regulation in many states represents a failure of the jurisdictions to evolve with the technology and markets the departments manage (Eling & Pankoke, 2016; Frezal & Barry, 2019; Schwarcz, 2018). Landes (2015) argued there are two conceptions in property and casualty insurance, one is having the ability to distinguish between insurance as a general association agreement among individuals based on risk pooling and specific agreements where a carrier acts as an intermediary between policyholders. The distinction provides different definitions of fairness where one is based on the collaboration between individuals in the risk pool, where the other is based on the equity of the contract between the insurer and the insured (Frezal & Barry, 2019). The second concept is premiums should be developed on the expected losses to better reflect the risk of the insured (Landes, 2015).

In property and casualty insurance, the ability to accurately rate and assign a premium to the individual risk is defined as actuarial fairness, which has evolved to be more powerful with the influx of data and advancements in data science modeling. Eling and Pankoke (2016) stated the appearance of data science magnifies old debates concerning the fairness of insurance ratemaking, with the introduction of new data supporting the consumer. The role of historical insurance regulations becomes even more unclear. The Consumer Federation of America (CFA), in its disapproval of price optimization and the fixing of margins based on the consumer's willingness to pay, stressed the movement of pricing away from the historical cost-based approach to be unfairly discriminatory (Frezal & Barry, 2019).

Insurance commissioners universally provide the guidelines where rates should be adequate to cover loss costs and are not considered discriminatory, as in the case of charging different rates for risks of a similar underwriting background. Premiums paid by consumers should match their overall risk exposure as closely as possible and are the products of the probabilities of losses and expected losses (Landes, 2015). Rate regulations in the auto insurance industry are designed to ensure the carrier rates are not excessive or unfairly discriminatory, leading to underpricing riskier consumers leaving less risky people to subsidize the earned premiums needed to cover incurred losses (Schwarcz, 2018). When requirements are not met, such as a carrier not validating underwriting criteria or rating variables, resources are transferred to the low-risk individuals to the high-risk individuals, and the insurance instrument changes the policyholders' expected value (Lehtonen & Liukko, 2015). The insurance approach is actuarially unfair because the expected benefit of the low-risk consumer is reduced, while the expected benefit of the high-risk consumer is increased.

The increasing complexity and number of regulations are also highlighted as being an important threat to the insurance industry, and there is little to no literature on the costs and benefits of insurance regulation due to the difficulty of capturing and measuring the information (Eling & Pankoke, 2016). Economic research suggests insurance rate regulation is not in the public interest as studies have documented the markets thrive when states can deregulate rules as the industry has several hundred competing carriers (Lehtonen & Liukko, 2015). In line with these findings, property and casualty carriers in most markets can earn reasonable and not excessive profit margins compared to other financial service industries. For the present time, rate regulation is often ineffective, cyclical, politicized, and can effectively weaken competitive market conditions by discouraging market entry and causing individual carriers to dissolve (Schwarcz, 2018).

Big Data and Regulation. The combination of technology, effects, and capabilities has been rolled up under the definition of big data (Zuboff, 2015). Big data tends to be an arrangement of information systems and the ability to separate vast quantities of data to allow data mining for patterns to be used for predictive analytics (Yeung, 2016; Zuboff, 2015). For the insurance industry, catching up with the tidal wave of additional available data has allowed machine learning algorithms on various datasets, which had been previously unavailable. Over the past 20 years, the United States and other countries have started to prohibit insurance discrimination based on group characteristics, so finding additional variables from new data sources is critical to the rate segmentation goals for insurance carriers (Meyers & Van Hoyweghen, 2018).

Modern data protection regulations somewhat rely on a model of privacy self-management where the law provides consumers with a set of rights, enables them to control their

personal data, understanding those people will weigh the benefits and costs of sharing their data (Yeung, 2016). Cohen (2015) argued the approach places the responsibility on both parties as the party creating or capturing the data must notify the consumer, and the consumer needs to consent to the actions taking place. Critics, including consumer advocates, argue the consumer is unlikely to give a tremendous amount of thought to consent to data sharing and are invited to share personal data in return to digital services and online purchases (Zuboff, 2015). There is overwhelming evidence the consumer neither reads nor understands the online privacy policies, and most companies try to enable practical solutions to provide informative notices and protect personal information (Cohen, 2015; Meyers & Van Hoyweghen, 2018; Yeung, 2016). For entering into a financial contract such as auto insurance, carriers must have the consumer sign an acknowledgment statement they understand a third-party entity will be providing their public and personal data, which will be used for underwriting and rating purposes. Departments of Insurance have increasingly taken cues from consumer advocates and legal groups trying to remove private data sources because they feel individuals are not making sound decisions when making complex decisions involving privacy in specific contexts (Zuboff, 2015). When new data sources are introduced from third-party vendors, the regulatory approval process is extensive and ensures the carriers are transparent concerning the data being ordered and the consumer understands what data are being used for ratemaking and underwriting for segmentation purposes.

Adverse Selection. It has become important for insurance carriers to innovate when designing *a priori* and *a posteriori* rating approach by developing representative probabilistic models based on statistical methods for distributing the number of claims allowing the carriers to match premiums to risks fairly and equitably. The integration of new types of pricing data and

verified rating variables allows carriers to attract and retain consumers by adopting innovations, which allow the carrier to lower premiums and improve products (Tzougas et al., 2019). If an insurance carrier can develop new pricing methods, or introduce new rating variables, to identify and attract lower-risk individuals, those individuals would have a financial incentive to change to the innovative carrier to save money by paying a lower rate to match their level of riskiness (Cather, 2018).

One carrier's gain is another carrier's loss, and the loss of lower-risk insureds can affect carriers' profitability by taking away the lower loss ratio portion of the portfolio through the adverse selection process. Those carriers which select not to innovate cannot survive in the highly competitive market place, and the hundreds of carriers in the market will continue to file new rate class plans with updated approaches and pricing structures. Demand friction, the economic friction keeping markets from operating according to the perfect competition model, affects the classification of individuals who have different valuations along the demand curve (Spinnewijn, 2017). Spinnewijn (2017) stated the studies analyzing the importance of adverse selection in insurance markets reveal the insurance contract's overall value and the importance of pricing each risk adequately.

The Future of Property and Casualty Insurance Data and Modeling. Nicholson (2019) stated a range of factors are shaping the future of the insurance industry, including technology and consumer expectations, which have become compelling forces based on recent trends and developments in a competitive environment. Technology allows for more innovation and efficiencies for newly formed insurance carriers, which are customer-centric and offer better products and services with new business models designed to disrupt the traditional property and casualty markets (Charpentier, 2017). The traditional insurance market is very slow to change,

and those carriers embracing the new data and statistical advancements are experiencing a shift away from traditional actuarial science to a robust data science structure for product pricing. Binder and Muβhoff (2017) of McKinsey & Company suggested in the next 10 years, 40 percent of the insurance industry positions, which currently exist, will no longer be needed, and 20 percent of the positions needed in the future do not exist in the industry today. Binder and Muβhoff (2017) also predicted the property and casualty insurance industry would change more in the upcoming years than it has in the past 100 years.

New open-source data and the advancements in statistical modeling in open code are also allowing rapid development in artificial intelligence and machine learning, which are creating products and services with the potential to continually disrupt the environment actuaries were used to studying (Richman, 2018). The future expansions of quantum computers and high-level quantum algorithms will ignite an explosion of modeling possibilities, likely to restructure the entire industry (Charpentier, 2017). With the transition to insurance digital platforms, insureds will have a portal potentially consisting of blockchain technology, artificial intelligence, computer learning, complex algorithms, and big data sources with insurance carriers at the end of the portal (Nicholson, 2019). With new verified third-party data, matching interests can be paired with optimal premiums where carriers could operate at high efficiency, speed, and accessibility.

Usage-Based Insurance. From an insurance rating standpoint, the more a vehicle is driven on the road, the larger the overall exposure period and the vehicle's time to be involved in a loss event. Traditional premium calculation approaches for auto insurance carriers were mainly based on general factors such as observable vehicle make factors and driver demographic variables, which could be collected and verified for rate orders of premium calculation (Casualty Actuarial Society, 1990). Most carriers supplement the *a priori* approach to premium

development with driver violations and accident history to improve the overall rate accuracy. Even though there was broad adoption and use of the base models created decades ago, there are still limitations, specifically with the actual vehicle exposure, which is challenging to capture and validate (Baecke & Bocca, 2017).

Pay-As-You-Drive. Consumers are charged directly for where and how are far they drive in Pay-As-You-Drive programs, which are possibly due to new information and communication technology (Dijksterhuis et al., 2015). The Pay-As-You-Drive programs addressed what consumers felt were shortcomings in traditional insurance programs as the programs are viewed as being transparent and fair because individuals are paying for coverage based on their own driving behavior instead of being charged a rate from an aggregated pool of consumers. Individuals who drive safely and defensively should pay lower premiums, and consumers who drive risky and aggressively should be paying higher premiums to cover their loss costs (Weidner et al., 2017).

The challenge with Pay-As-You-Drive programs is validating even the most straightforward data, such as data and mileage from odometer readings, without the cost of monitoring outstripping the benefit of collecting the data. Self-reporting is not an adequate measurement as individuals have no incentive to provide accurate information. The insured needs to be willing to have the information verified with digital information capture and could range from pre-paid premiums by purchasing miles to full behavior-based telematics Pay-As-You-Drive insurance (Dijksterhuis et al., 2015). With more people enrolling in such programs, benefits from large-scale usage will start to drive down device costs as already experienced with digital data (Husnjak et al., 2015).

Telematics. Devices installed into vehicles through the onboard diagnostics port can interface and collect data for both driver behavior and actual miles driven validating the real-time exposures for the insurance carrier (Husnjak et al., 2015). The original data collection devices needed to have the data pushed to the insurance carriers and included all of the data the device was equipped to store and transmit. From a consumer perspective, connecting the miles driven to the vehicle would be the most accurate way to assess exposures. The driving behavior variables can also be evaluated and used for premium calculations, and univariate predictive performance over three months is long enough to obtain enough data with the highest predictive power (Baecke & Bocca, 2017). Driving behavior over time follows the patterns of normal behavior and is relatively consistent. If driving data from over one to four years is used, some individual driving behavior is not current enough to be considered for predictive modeling (Baecke & Bocca, 2017).

Large amounts of data and digital technologies have raised privacy concerns, and security issues as the information being transferred between electronic devices compare to a form of involuntary user observation without providing disclosure statements (Barth & de Jong, 2017). With the increase in data sharing across digital platforms, the consumers need to find and understand how the data are being used, and ultimately, which entity owns the data being collected. An individual's driving habits include the time of day, patterns, hard braking, distance traveled, and of most interest to the insurance carrier, speed, and acceleration (Bellatti et al., 2017; Vavouranakis et al., 2017). Surveys report the consumers are open to participating in telematics programs, but a significant portion of drivers are concerned the insurance carriers may be tracking their destinations and sharing data with unknown third parties (Bellatti et al., 2017). While carriers are careful about not sharing personal identification indicators, the possibility of

linking via quasi-indicators remains. Research from online service providers shows reliable privacy statements and risk awareness do not change in line with modified preferences, and although users are aware of privacy risks, they tend to share private information in exchange for discounts and personalized services (Baecke & Bocca, 2017).

Mobile Applications. While providing some of the most predictive variables for insurance premium development, onboard diagnostics telematics can be an expensive method for collecting verified data for insurance carriers who specialize in financial services. Innovations in smartphones and mobile devices include advanced motions sensors and accelerometers providing, a new platform of distributed sensing devices (Castignani et al., 2015). Insurance carriers are working with vendors who can pair the driver to the vehicle and collect mileage and driving behavior data. Vavouranakis et al. (2017) agreed the best practice for recognizing driving patterns is using the accelerometer data from the smartphone technology due to the ability of the algorithm to detect sharp turns, lane changes, and increased speed.

The drawbacks of mobile device technology are there are human decisions involved, and the smartphone device needs to be with the driver for device-to-vehicle pairing. The different input variables are collected at different rates, and an interpretation layer needs to be introduced to perform event detection based on different time series from different devices and data carriers along with normalization and scoring logic (Castignani et al., 2015). Mobile technology can help provide aggregated benefits, including changing the risk assessment process, including the ratemaking process, which results in lowering overall risk and obtaining dynamic statistics and verified driving behavior data (Ohlsson et al., 2015). The challenge is getting consumers to actively download insurance applications and allow carriers access to private information to determine a premium (Vavouranakis et al., 2017).

Modeling Approaches

Insurance carriers depend on statistical models to predict future claims events and forecast financial protection for both the consumer and the company (Davoudi Kakhki et al., 2018). There are also requirements set by the independent state departments of insurance to provide actuarial, or mathematical and statistical exhibits to ensure the right methodology, approaches, and outcomes are used to set rates within each program. Understanding the dependent relationships between a dataset of predictor variables and zero-inflated count outcomes, which follow the traditional insurance claims patterns, are the basis for comprehending risk factors driving the ratemaking decisions (Chowdhury et al., 2019). The zero-inflation occurrences are because the carriers only reimburse for losses, which meet the contract parameters or exceed a threshold. There is also evidence individual insureds would rather pay for low payment loss events to avoid a potential premium increase for reporting the claim and receiving reimbursement (Davoudi Kakhki et al., 2018).

Multiple Linear Regression. Regression analysis, or linear modeling, is a core approach in statistical modeling used for auto insurance data analysis and decision support (Fox, 2016). Of particular use for auto insurance modeling, linear models have three different uses of summarizing data, predicting future events, and predicting the results of interventions (Frees et al., 2014). In predictive modeling, carriers analyze data to build models, which will be used to estimate an unknown future quantity using one or more known independent predictor variables. Fox (2016) stated for linear regression approaches, analysts are interested in using characteristics of policyholders such as age, sex, and driving experience along with the vehicle attributes to help explain the target-dependent variable, pure premium. Modeling insurance risks with high accuracy and prediction rates is critical to insurance carriers as the modeling of past claims is

necessary for estimating loss costs, which is the cornerstone of pricing in the insurance industry (Davoudi Kakhki et al., 2018).

Classification is used in the property and casualty insurance industry based on attributes of the rating variables, and the performance of the classification process is dependent on how well the discriminant function, which is a function of variables used to assign independent variables into one of two or more groups (Khashei et al., 2012). The classification process is used to minimize the misclassification rate for the targeted problems performed on both the modeling data set and the hold out data set. The classification approach of assigning numerous events into different separate groups performs a critical role in business decision-making for insurance ratemaking (Davoudi Kakhki et al., 2018). Combining several models or creating hybrid models improves predictive performance using multiple linear regression models to return additional general and more accurate models (Khashei et al., 2012).

Generalized Linear Modeling. Average losses are expected to equate to pure premium, which is the product of expected claims frequency and severity, resulting in at least a financial breakeven state. Statistical approaches can be used with frequency and severity models to classify the risk classes characterized by independent rating factors (Xie & Lawniczak, 2018). Insurance carriers use Generalized Linear Modeling for risk classifications and can allow for response errors to develop from exponential outcomes from modeling. There has been much research and advancements for predictive modeling when finding the optimal solution to reduce overall bias. Generalized Linear Models possess the property of providing unbiased estimates on the book of business level, which suggests the models can also deliver accurate prices at the business level (Wüthrich, 2019). Carriers will model loss costs, and the bias is defined as the difference in the predicted loss costs and the actual observed value (Fuzi et al., 2016). A

minimum bias method is an approach used by insurance carriers to assign restrictions within classification ratemaking. Fuzi et al. (2016) stated the minimum overall bias considers both the estimation of premium, the product of frequency and severity for each class of insured, and the related number of earned exposures.

The Generalized Linear Models are considered acceptable approaches for the department of insurance filings because the models can be described and replicated by the carriers and other entities reviewing the filings for adequate actuarial support. Insurance regulators are not interested in reproducing the carrier models or results. Insurance departments need an analysis based on the aggregated loss experience for their reviews and analysis (Xie & Lawniczak, 2018). A filing review conducted by a regulator needs to be supported by key findings based on the rating factors used in developing the model and at the industry level. However, the models also have drawbacks, which require the modeler to specify how the covariates interact and what functional form they should take in the resulting regression function (Wüthrich, 2019). Insurance carriers use several dozen risk factors for rating, which also raises the question of whether a carrier or the regulators can focus only on the most predictive rating factors during the approval process (Xie & Lawniczak, 2018).

Interaction Variables. An issue, which may need to be addressed in modeling claims frequency and severity, is the modeling of interactions (Valecky, 2016). The data preparations for multiple linear regression or generalized linear modeling will focus on optimizing the selection and transformation of variables with the assumption the variables in the dataset have an individual effect on the dependent variable (Goldburd et al., 2020). Researchers need to consider the situation where two or more variables may have a related effect on the dependent variable, which creates a situation that could have an inadvertent inflated outcome on the target variables.

Goldburd et al. (2020) called the effect and dependency of one predictor variable on another, an interaction. In auto insurance claims modeling, a potential interaction set of variables could be *age x gender*, which can be added to the pure premium model to determine if the model outcomes are affected (Valecky, 2016).

Tweedie Compound Poisson Models. Claims events in insurance are considered to be an all or nothing proposition. For insurance claims models, loss events occur in a probability mass at zero or in a non-negative amount higher than zero over time, following a right-skewed Tweedie compound Poisson distribution (Qian et al., 2016). The challenge for the insurance carriers is to predict the size and rate of recurrence of future claims accurately. Yang et al. (2018) stated highly right-skewed, mixed point data with a point mass at zero cannot use transformation techniques to normality by power transformation, and will employ Tobit models or Tweedie distributions to simultaneously model frequency and severity of claims. Due to the Tweedie Generalized Linear Model's ability to model zeros and continuous positive outcomes simultaneously, the approach is a highly used for auto actuarial study methodology. The most significant drawback of the Tweedie Generalized Linear Model is the underlying logarithmic mean, which is a linear shape and not flexible enough for auto insurance modeling (Yang et al., 2018). Qian et al. (2016) noted in auto insurance modeling, the risk does not monotonically decrease as age increases, although nonlinearity can be modeled by adding splines. Low-degree splines may not be suitable to capture the nonlinearity and high-degree splines often result in overfitting, which in turn produce unstable estimates (Qian et al., 2016). To reduce the restrictive linear assumptions of Generalized Linear Models, Generalized Additive Models can model the continuous variables by smoothing functions estimated from the data (Yang et al., 2018).

Advanced Statistical Pricing Models. Moro et al. (2017) stated the increase of knowledge through data mining had driven carriers to increase their accumulation of data, but there is a gap between having the ability to access more information and the application of knowledge and improved decision-making. For many insurance carriers, having more data does not necessarily indicate better predictive pricing models. The data have provided the opportunity to implement flexible models with machine learning techniques becoming more mainstream in data analytics, providing highly configurable and accurate algorithms, which can work well with both structured and unstructured data (Baudry & Robert, 2019).

The use of decision trees has become a progressively accepted alternative predictive modeling tool for building classification and regression models (Quan & Valdez, 2018). Carriers can increase premiums and filter risks through underwriting rules to balance the underlying risks (Bärtl & Krummaker, 2020). Innovations to the original approaches, such as random forests and gradient boosting models, have broadened the capabilities of using decision trees as predictive models (Henckaerts et al., 2018). Decision trees using multivariate response variables can model correlated responses when using insurance claims data. The addition of multivariate response variables provides advantages of the univariate decision tree models to rank critical explanatory variables and high predictive accuracy (Quan & Valdez, 2018).

While advanced statistical modeling is applicable and relevant for predictive modeling in auto insurance, the respective insurance departments will only allow rate-filing submissions with multivariate regression models or Generalized Linear Models. The data and technological advancements have steered data scientists toward machine learning and big data analytics, which changed the trajectory of predictive analytics and statistical modeling. Few peer-reviewed papers in property and casualty insurance literature go beyond Generalized Linear Models (Henckaerts

et al., 2018). Actuaries and product managers need transparent and interpretable ways to explain pricing models to stakeholders and the insurance departments. Insurance pricing models are heavily regulated and need to meet specific actuarial support requirements, which establish a means for algorithmic accountability (Pasquale, 2015). Kaminski (2019) argued consumers have the right to understand the logic behind the pricing decisions, which are used to determine the premium for coverage. Pasquale (2015) stated the consumer, product managers, and regulators should receive the information in varying levels of detail and scope, and the insured should be charged a fair premium base on their risk profile to reduce adverse selection. If the diversity of the portfolio is not reflected in the ratemaking, lower-risk insureds will non-renew their coverage and would instead seek coverage, which priced the premium based on the merits of the individual's risk profile (Wüthrich & Buser, 2019). By finding balance in consumer segmentation and risk pooling, the insurance carrier can avoid adverse selection and offer the insured the right rate.

Price Optimization. In mature markets, increases in data and modeling sophistication allow carriers to optimize pricing approaches (Spedicato et al., 2018). Insurance carriers are looking to model the cost of the insurance coverage and the overall demand for a commodity product. The loss cost approaches to pricing individual risks define the price of a policy as the ratio all future claims' estimated costs compared to the coverage provided to cover the exposure (Guelman & Guillén, 2014). Indicated rate changes project a revised set of rates, which will directly impact the profitability of a book of business and the amount of premium charged and the insured's reaction to the rate change. Spedicato et al. (2018) argued it is now to an insurance carrier's advantage to model the cost of the coverage offered and the overall demand before and

after rate changes. Current policyholder retention modeling, and with consumer conversion probability modeling provide the basis for pricing optimization (Bolancé et al., 2018).

The value of insurance consumers is higher in those who create larger profit margins, which are greatly influenced by price and highly dependent on the consumer's willingness to purchase the policy (Guelman & Guillén, 2014). Price reduction models-based optimization modeling would be useful and would require a dynamic way of setting prices for auto insurance, which is against state regulations requiring the filing and approval of rating structures. Bolancé et al. (2018) acknowledged the current regulatory environment lags the necessary flexibility to allow for proactive price optimization and allowing for consumer pricing tiers based in part on observed and historical demand behavior.

Variables in the Study

The existing pricing methods and approaches for auto insurance ratemaking are evolving with increased involvement due to changes in the political and economic environments and the changing consumer needs from the industry (Störmer, 2015). Bernard et al. (2015) argued the consumer faced an unpredictable random loss and could select to pay their insurance carrier a premium in return for sharing the risk of future loss. The ideal design involves determining the amount of loss covered by the insurance carrier, which is defined as the indemnity, and forecasting the corresponding premium covering both the carrier and the insured's security (Bernard et al., 2015). The changing consumer needs also shape the pricing structure. Seventy-five percent of insurance carriers have a pricing focus on risk-based costs over consumer-based pricing (Störmer, 2015). Carriers trying to generate profitable and equitable growth, strive to strike a balance with validated rating variables, which aligns better with the overall individual risk and customer-oriented ratemaking.

Bonus-Malus System. In auto insurance, *a posteriori* rating used for premium development is used to compliment *a priori* rating and consists of Bonus-Malus System levels, which build on each other to create the ratemaking environment (Tan et al., 2015). The carrier determines the Bonus-Malus System levels, and each consumer enters at a starting level and renewal, moves up or down the levels based on transaction rules depending on the risk's claims and driver behavior changes during the policy period (Tan, 2016b). Insurance carriers will continue to monitor all drivers and vehicles on the policy for change created by claims activity on all vehicles and violations incurred by each driver. The levels have premium adjustment coefficient relativities assigned to independent rating variables multiplied with a base premium in a rate order of calculation to determine the final premium amount (Tan et al., 2015). Drivers are assigned to a vehicle along with their associated violations, or an average of all policy violations for average driver rating to calculate vehicle level premiums, which are aggregated in the rate order of calculation to determine the final policy premium.

Risk-Based Pricing Model. In a risk-based auto insurance pricing model, the rates charged to individual insureds are customized to reflect the underlying risk characteristics and driving behavior of the combination of vehicles and drivers listed on the policy (Isotupa et al., 2019). One of the critical items to note is auto insurance carriers will only cover loss events on listed vehicles on the policy and can only accurately price the entire risk if all drivers with access to the vehicles are also listed and rated. Undeclared drivers are a known risk, and carriers will address the underpriced insurance fraud from not listing the underlying risks with various forms of third-party underwriting reports for household members. In a social pricing model, all policyholders will pay the same premium in an average pool-based rating approach (Störmer, 2015). With risk-based pricing approaches, higher-risk insureds should pay higher premiums

than low-risk insureds as the vehicle and *a priori*, observable characteristics, are correlated with the anticipated loss experience (Isotupa et al., 2019). The insurance departments in the United States use risk-based models. The regulations vary significantly on whether or not uncontrollable characteristics such as age, gender, and marital status can be used as rating variables, or only allow those characteristics such as type of car, garaging zip code, and the distance each vehicle is driven.

Rate Order of Calculation. The rating algorithm used to develop premiums for each risk needs to consider the various independent rating variables in conjunction with the underwriting rules. There is a separate base rate for each line coverage, which is not an average rate, and the coefficient factors are applied as either a series of multipliers, addends, or some unique mathematical equation as defined in the rate order of calculation (Werner & Modlin, 2016). The complexity of the rate order of calculation, and the order of the rating variables, will vary by the insurance carrier. Those carriers who can identify predictive independent variables outside of the mainstream rating variables (e.g., credit score, prior bodily injury limits, and homeownership) will differentiate themselves in the marketplace, allowing for a more refined risk-based insurance model with rating characteristics, which match the individual risk profile.

Pure Premium. Werner and Modlin (2016) defined pure premium, or loss costs, as the measurement of average losses per exposure unit and described the portion of the risk's expected costs, which were purely aligned to losses:

$$Pure\ Premium = \frac{Losses}{Number\ of\ Exposures}\ Frequency\ x\ Severity$$

The assessment of pure premiums has highlighted an essential issue in modeling and estimating of risk classification factors, by selecting the appropriate approach to pure premium modeling

(Jee, 1989). A central issue arising from estimating pure premium, or claims frequency and severity, is selecting the appropriate form for the statistical model. Using a flexible category such as a hybrid model would take advantage of both additive and multiplicative functional forms for estimating the pure premium. The estimation of the pure premiums for rate class plans using regression approaches requires a function form for the statistical model with the most customary options being linear and log-linear models (Harrington, 1986). There is also a need for refined statistical modeling and methods for selecting interaction terms for pure premium models to improve overall predictive accuracy and mitigate bias in predicted values (Jee, 1989).

Claims Frequency and Severity Modeling. The modeling of auto insurance claims frequency and severity is a detailed and important process providing the foundation for ratemaking for the various line coverages (Park et al., 2018). The claims experience for auto losses is conditional on random outcomes of claims frequencies and severities. By design, a risk may incur and more than one claim can happen during the policy period and the amount on any given portfolio is challenging to predict (Omari et al., 2018). Insurance carriers need to settle claims for future periods, and it is critical for the insurers to sufficiently model historical and current claims data to forecast future claims for ratemaking and reserving. Park et al. (2018) stated carriers are confronted with the challenges of selecting applicable statistical distributions for claims data and determining how well the designated distributions fit. Claims frequency and severity are often dependent, with claims counts being negatively associated with collision coverages because drivers who file multiple claims per policy term are typically involved in minor loss events (Garrido et al., 2016).

There are two central approaches for combined claims modeling to account for the potential dependency between the claims frequency and severity. Shi et al. (2015) compared the

conditioning and use of claims count as a covariate for modeling the average claim size distribution and marginal Generalized Linear Models to fit the model to the frequency and severity elements and link them through a copula, which is a function that couples multivariate distribution functions to one-dimensional marginal distribution functions. The most common model used in modern auto insurance ratemaking is a frequency-severity model based on a Generalized Linear Model in which the claims frequency is a covariate in the severity regression model (Frees et al., 2014; Garrido et al., 2016; Jeong et al., 2017).

Insurance-Based Credit Score. Act 1452 of 2003, is referred to as the National Conference of Insurance Legislators (NCOIL) Model, which was developed to integrate insurance-based credit scoring for personal auto lines of insurance (Lacy, 2017). The Act addresses insurance-based credit for underwriting and ratemaking (Insurance Information Institute, 2019). The Act also prohibits certain events from being considered under the Federal Fair Credit Reporting Act (FCRA), which requires notices to individuals applying for insurance and ensures specific actuarial justifications to be met before a carrier can use insurance-based credit for underwriting or ratemaking (Lacy, 2017).

Insurance carriers use insurance-based credit scores with the actuarial support that how an individual manages their finances is a good predictor of loss activity; statistically, people with a lower insurance-based credit score are more likely to file a claim (Brockett & Golden, 2007; Insurance Information Institute, 2019; Lee et al., 2005; Stiff et al., 2019). Adding the insurance-based credit score as a rating variable from either proprietary or credit vendor developed models allows the insurance carrier to pair the risk with an accurate premium helping to prevent lower-risk drivers from subsidizing higher-risk drivers. The Insurance Information Institute (2019) highlighted it is important to recognize insurance-based credit scores are not the same as

consumer-credit scores as one score predicts insurance losses, and the other predicts credit delinquencies. Insurance carriers can also order insurance-based credit reports on the primary named insured on a policy without an extensive list of personal information (Einav et al., 2016).

Data such as ethnic group, income, age, gender, address, and marital status are not part of the insurance-based credit score, yet state departments of insurance still feel correlations are being developed based on suspect classifications (Stiff et al., 2019). Andreeva and Matuszyk (2019) argued the insurance industry expressed concern, since gender was associated with risk, of removing the variable would lead to higher insurance premiums for women. Regulators make assumptions that by removing a variable from the overall insurance rate order of calculation, the individuals' premiums in the corresponding risk classes would be the same. However, differences will remain due to proxy variables, which are approved for use in underwriting and rating. Regulators recognize the correlation of the uncertain variables and the resulting insensitivity of the model's predictive strength, yet there is no course of action for how those correlations should be addressed and the overall suggested path to a solution (Einav et al., 2016).

An insurance-based credit score will use information from a driver's credit history and confirmed to be statistically correlated with loss costs (Insurance Information Institute, 2019; Liu & Schumann, 2005). The various state departments of insurance will restrict some factors within the insurance-based credit score, such as total available credit, debt from financing for health reasons, and certain types of credit lines (Lacy, 2017). The introduction of big data analytics also leads to carriers having new information such as the underlying insurance-based credit variables and improved computing power, which opened up advanced predictive analytics. Carriers have data and advanced statistics to work with daily, which increased concerns about the lingering risk of discrimination for certain risk classes (Favaretto et al., 2019).

Liu and Schumann (2005) stated the challenges carriers encounter when using classification algorithms to build insurance-based scoring models are selecting the variables from several private and public data sources. The variables used for modeling may have many observations and a large number of associated features, which may be unrelated to the insurance-based credit predictive strength, but also may be redundant due to their high intercorrelation (Žliobaitė, 2017). Data Science teams in the industry spend time after the data preparation running correlation analysis to quantify the linear association's direction and strength between two variables. Without the intermediate correlation analysis, the classification algorithms would deteriorate from having several irrelevant and redundant features, which would increase the computation time, and decrease the model accuracy and scoring interpretation (Liu & Schumann, 2005).

Criminal Background. The insurance industry has not addressed individual criminal background in part due to the inability to verify a person's history, and there is also a concern about the perception of potential bias in criminal history data. Carriers are progressively relying on data-driven predictive models to help with decision-making, and due to the disruptive nature of the observational data, the models may systematically disadvantage individuals belonging to specific risk groups (Žliobaitė, 2017). The occurrence of systematically disadvantaging a driver in an auto insurance risk group may happen even if all personal identification information is removed before modeling. As the insurance industry begins taking greater advantage of data analytics and additional available personal data, socially sensitive decisions could have weighty impacts on individuals such as higher rates for auto insurance, and in some cases, a refusal for coverage (Carmichael et al., 2016). As carriers move toward more multifaceted forms of socially sensitive decision-making, the analytical groups need to find ways to ensure the use of data

mined and used for modeling does not lead to unjust discrimination against certain groups due to age, ethnicity, or gender (Roy & George, 2017).

One area of criminal background, which remains in the forefront of insurance investigations, is the problem with insurance fraud globally due to its low-risk, high-reward payoff attracting criminals and creating a financial burden to carriers and policyholders (Derring et al., 2006; Nagrecha et al., 2018). The fraud hit rates tend to be higher for non-standard auto carriers who write drivers who fall outside of the lower-risk category and have trouble paying insurance premiums. Nagrecha et al. (2018) argued some individuals who pose a higher risk to the insurance carriers are those drivers who incur multiple accidents, have prior violation convictions, and prefer to carry state minimum liability insurance coverage. By identifying these individuals before binding coverage, the insurance carriers can properly underwrite, or correctly price, the risks and reduce rate subsidization for lower-risk insureds (Barraclough et al., 2016).

Nagrecha et al.'s (2018) study referenced the association of higher-risk individuals and those consumers who are more likely to commit insurance fraud. However, the literature is silent on the claim frequency and severity and the predictive strength of overall criminal behavior. Until recently, insurance carriers could ask a misdemeanor or felony conviction question but could not verify the answers. Third-party vendors can now supply this information and determine if there is a correlation between loss costs and criminal behavior. Crimes are offenses against society, and individuals have been observed committing crimes at any time and place and in any form (Tayal et al., 2015). Society is careful about assessing penalties to crimes to ensure people are not continuously reprimanded for the same crime in several different areas. In societal contexts, the rationale for committing crimes appears to correlate with other risky decision-making aspects by individuals (Lipton, 2018). The larger question becomes, can predictive

modeling support the correlation between criminal behavior and increased in auto insurance loss costs.

Motor Vehicle Moving Violations. Between 1899 and 1903, hundreds of state and local governments passed legislation to address motor vehicle accidents and had little coordination between jurisdictions and public officials (Vinsel, 2019). The current regulations continue to be disjointed and governed by state and local law enforcement agencies. Each state defines and codes basic violations differently for each of the 50 states. Violations are grouped in categories, which closely follow the classifications in the vehicle code for each state, with only moving violations being considered for rating evaluation and non-moving and clerical citations being removed from the analysis (Palumbo et al., 2019). In the case of multiple violations occurring on the same day, the violation and conviction with the highest associated rating violation are assigned for rating purposes.

The driver violation calculations will use a driver-averaging rating approach, which calculates an average driver relativity based on the driver-violation-accident combination. Instead of using the assigned individual driver relativities, the approach uses an average of all of the drivers on the policy. The approach minimizes questions about the highest driver rating to the highest vehicle rating as well and prevents assignment questions mitigating concerns about riskier driver manipulation (Blesa et al., 2020).

Moving Violations and Criminal Background. One area of agreement between the respective departments of insurance is treating and counting driver moving violations and criminal driver background. While carriers are permitted to account for convictions incurred by driving and risky personal behavior, the carrier cannot double-charge for a conviction if it could potentially occur in both categories (Feest, 1968). The data preparation for criminal background

includes careful scrubbing to review conviction data types to ensure two violations do not appear on different reports. The data sources are also different as moving violations originate from the traffic court system, and the state departments of motor vehicles and the criminal driver background occurrences are obtained from the state and federal departments of correction.

The data source for the research study has both the datasets containing violations and convictions from both the departments of motor vehicles and the state and federal departments of correction. Data are cross-referenced to ensure a criminal conviction such as vehicular homicide is categorized and validated through the jurisdiction where the incident occurred. The departments of insurance also treat the categories differently. Insurance carriers can only use between 24 and 36 months of driving moving convictions for underwriting and rating, where carriers can use between eight to ten years of criminal driver background data for evaluation (Vinsel, 2019).

Historical Claims. Carriers need to settle loss events, which may occur from in-force books of business in future periods, and the historical claims experience can be used to model current and future events and to set reserves (Omari et al., 2018). Claims frequency is a predictor of driver risk levels and depends on both the individual's claims history and the current evaluation period (Asamoah, 2016). Miljkovic and Grün (2016) argued the modeling of insurance claims events tends to be more forecasting art than actuarial science. A modeler will need to analyze and make well-supported assumptions during the modeling process for pricing and reserving.

The data sets containing claims data have many zeros, no payment incurred, which provides researchers with a fat-tailed, right-skewed distribution rather than a normally distributed set of data (Gan & Valdez, 2018). An error in the modeling on one end of the tail can negatively

influence premium and loss reserve adequacy. For the study, data transformation will involve creating a new distribution based on the logarithmic transformation of the claims' amount to create a symmetric distribution for modeling purposes (Miljkovic & Grün, 2016). The generalized linear model approach is based on the normal distribution of modeling approaches, and the distributions can be discrete, continuous, or both combined (Asamoah, 2016).

Proof of Prior Insurance Coverage. Proof of prior insurance coverage verifies a consumer carried in-force state minimum limits liability coverages for vehicles without a lapse between coverage effective dates. For traditional standard auto insurance markets, insureds with prior coverage allowed carriers to offer discounted rates as carriers acknowledge there is a robust unconditional relationship between policyholder tenure and the policy's overall riskiness (Kofman & Nini, 2013). While some researchers argue about average switching costs, which are substantial when comparing acquisition expenses, the higher retention rates for standard insureds center on customer claim satisfaction during the policy period (Honka, 2014). Insureds tend to look for another insurance carrier when they have had a claim, and the process, coverage, or payment did not meet expectations (Barraclough et al., 2016).

The non-standard insurance market is defined differently by the various property and casualty insurance carriers and ranges from higher-risk behavior with multiple driver violations to frequent accidents or state minimum liability coverage limits, and most importantly, the inability to pay insurance premiums. Proof of prior insurance was one of the first proxies for financial responsibility before insurance-based credit scores became one of the most robust predictive variables for estimating loss costs (Stiff et al., 2019). Lapses in prior insurance coverage can indicate higher than average claims frequency and severity for risk, and premium discounts can be offered to those insureds who persist longer (Honka, 2014). Carriers want to

attract lower-risk consumers to decrease portfolio loss costs and new business acquisition expenses.

The research study data will include auto insurance records with continuous insurance coverage and various levels of lapses in coverage. Kofman and Nini (2013) observed strong negative relationships between loss ratios, claims costs to premium, and policy persistency. Guelman and Guillén (2014) suggested insurance carriers would proactively reduce their profit margins slightly to increase overall renewal rates for a book of business. Bolancé et al. (2018) also argued the optimal pricing structure for insureds should consider the impact of the renewal book on the pricing segmentation for the overall portfolio of risks.

Driver Age and Experience. In terms of non-controllable, individual rating attributes, age has a high impact on rated premium development (Kelly & Nielson, 2006). Youthful drivers are new drivers who have recently earned a driving license and generally have not acquired enough driving experience compared to mature drivers with fully developed visual scanning strategies, which come from years of driving experience (Yeung & Wong, 2015). Age-related changes in perception and performance posed another important consideration at the other end of the age spectrum. Senior citizens comprise the fastest growing driving population and maturing adults are also associated with higher rates of collisions and fatalities as people experience a general decline of behavioral and cognitive abilities with age (Mikoski et al., 2019).

The age-value function over a life cycle has an inverted U-shape due to two patterns of marginal adverse events for cost and risk over a longer period (O'Brien, 2018). Seniors and middle-aged drivers have a comparable disutility from cost, and seniors place less importance on safety, resulting in a lower value of statistical life, which is a dollar value assigned to the benefits of health and safety regulations (O'Brien, 2018). Youthful drivers also place a lower weight on

safety and higher importance on the cost compared to middle-aged drivers, which suggests a lower value of statistical life for youthful drivers (Yeung & Wong, 2015). From a safety aspect, driver reaction time is critical, especially in complex, compact urban environments (Brown et al., 2007).

For the study, age groups are broken into logical categories and tend to align closely with other carriers writing similar target markets of risks. The actuarial analysis does support both youthful and mature drivers, classified into risk groups, present a greater risk for ratemaking purposes as a higher number of drivers in these groups will incur losses relative to other age group cohorts (Yeung & Wong, 2015). Insurance regulators have been under substantial pressure to ban the use of rating segmentation based on age, and consumers demand property and casualty carriers should not be able to discriminate. The ethical argument is the use of age is an individual characteristic variable which is discriminatory for risk classification and was deemed to be weak because an individual would face an equal probability of moving through each age classification group through the course of a driving lifetime (Brown et al., 2007).

Vehicle Rating. Many of the rating variable discussions centered on the various driver variable characteristics. The insurance carrier is providing traditional coverage for the listed vehicles on the policy, and those attributes are also carefully researched. Auto carriers have been aware of increasing vehicle repair costs, and as more information becomes available, insurers become more involved in the development and application of the data (Cole et al., 2015). The Highway Loss Data Institute collects and distributes data on collision claims frequency per 100 earned vehicle year for each model of vehicle, which is made available to the public and is used by the majority of property and casualty insurance carriers to help determine rating relativity for each make, model and year available (Joksch, 1981). The Highway Loss Data Institute

publication is still the gold standard in modeling make, model, body style, deductibles, and if there is, a youthful driver involved in the loss event. Cole et al. (2015) stated there are logical differences between the frequencies for vehicles with low versus high deductibles and with the inclusion of a youthful driver, so the frequencies need to be standardized to the same mix of deductibles, and driver age groups when modeling.

Of the 24 significant variables linked to the risk, the most relevant variables include: the age of the driver and the years of driving experience, accidents and violations occurring in the last five years, the type of vehicle, the vehicle power, the vehicle price, and the age of the vehicle (Blesa et al., 2020). The Insurance Services Office (ISO) provides auto insurance carriers with auto rating symbols, which define categories and rating relativities for vehicles based on year, make, and model, or the carrier develops a proprietary classification system for vehicle rating. The insurance carrier has developed a Vehicle Attribute Rating approach, which uses a Vehicle Symbol Rating generalized linear model based on historical policy and loss experience to establish if any vehicle attributes like curbweight, maximum horsepower, or base price significantly influence claims loss costs. A data set from the carrier used private passenger auto data from accident years 2004-2006 and was scrubbed for accurate records and tested for homogeneity. The policy accounts for the model data and contains the vehicle attributes from each Vehicle Identification Number (VIN) provided by the Highway Loss Data Institute. For pre-1981 vehicles, attributes are not available, and a separate analysis based on historical data was completed to develop a common symbol category factor.

Coverages, Limits, and Deductibles. Auto insurance covers damage to the insureds vehicle and includes damages to the driver and others due to an injury caused by an accident, such as medical treatments, payments for pain and suffering, and lost wages (Kadiyala &

Heaton, 2017). Liability insurance includes the coverages, which provide compensation for injuries. Depending on the type of policy purchased, it will provide compensation for either third-party or first-party damages up to the limits of liability. The Physical Damage coverages are those coverages, which protect the vehicles listed on the policy. Collision coverage will provide protection if the vehicle is involved in an accident with another vehicle or an object. Comprehensive coverage will protect for listed events incurred by the vehicle, excluding the events identified under the collision coverage (Gaffney & Ben-Israel, 2016).

For traditional consumers, purchasing insurance must increase the expected cost to the insured, or otherwise, the insurance carriers would not be able to remain solvent as the expected cost increase with higher limits or lower deductibles should be compensated with a decrease in risk (Kadiyala & Heaton, 2017). Most state governments, through the independent departments of insurance, require vehicle owners to carry state minimum liability limits to create a financial responsibility contact in case of an accident involving other drivers or passengers (Hsu et al., 2015). One item to note is the purchase of auto liability or physical damage coverages neither reduces the frequency or severity of accidents, and it only serves to mitigate the financial losses of individuals involved in a vehicle loss event (You & Li, 2017). The relationship in insurance markets is described by adverse selection theory, which describes situations in which insureds retain some of the information involving their overall risk level, which is unavailable to the insurance carriers when making underwriting or pricing decisions (Hsu et al., 2015). If adverse selection is existent, those drivers who purchase higher liability limits and lower physical damage deductibles are likely to be at higher risk in accident situations, creating a positive correlation between coverage and risk (You & Li, 2017).

For the research study, the data set will be auto insurance records with liability coverages purchased at state minimum financial responsibility limits and with collision and comprehensive coverages at five hundred dollars. The consistency between limits and deductibles will allow the researcher to have a standardized data set without needed additional transformations. Hsu et al. (2015) argued insured individuals who submit claims to their insurance carriers are more likely to have purchased higher limits and lower deductibles, so leveling the data set at state minimum limits will also standardize the claims frequency and severity distributions.

Summary of the Literature Review

For the auto insurance industry, carriers must analyze and select a highly predictive set of risk variables with their associated factors to predict accurate future claims distributions. The industry is experiencing an increase in validated data sources and benefiting from advanced computing power, allowing for the development and implementation of sophisticated pricing model approaches. The data-driven analytics allows auto insurance carriers to segment risks based on individual behavior characteristics to avoid average rating for a large pool of risks. Finding predictive variables, which can be verified and supported with actuarially sound statistical models, benefits the consumers purchasing the coverages and the carriers providing the financial instruments to distribute risk equitably.

Peer-reviewed articles from academic journals and primary source books were reviewed for actuarial ratemaking, variable selection, and predictive modeling for premium development and loss analysis. Journal articles focused on criminal insurance activity were directly addressing insurance fraud and the submission of fraudulent claims. Due to the lack of verifiable data, the literature is silent on criminal background behavior being collected or used to develop auto insurance premiums for risk segmentation.

Predictive models are used to segment insurance risks based on forecasted loss events using the underlying rating variables and historical claims data. Criminal behavior as a rating variable should be an associated high loss ratio, which carriers can show to be actionable. Instead of pricing the entire book of business with an assumed even probability of criminal background, an auto insurance carrier needs to consider introducing the additional variable to remove assumptions from premium development approaches. Due to state regulatory restrictions, rating variables must be observable and actuarially validated before being used for pricing segmentation. Until recently, property and casualty insurance carriers did not have substantiated individual criminal background data available for predictive modeling. The research study will not predict which drivers are more likely to commit criminal offenses, but will be able to determine if drivers with criminal backgrounds are unprofitable enough to warrant premium increases between 10 and 20 percent to offset higher loss costs.

Transition and Summary of Section 1

Research and the insurance industry have indicated there is a need for additional predictive variables to support further risk segmentation to avoid average pricing and rate subsidization by lower risk drivers (Ishida et al., 2016; Nagrecha et al., 2018; Nicholson, 2019; Roy & George, 2017). The research study will increase the depth of literature on predictive rating variables available for actuarial analysis and support for rate and rules filings at the individual departments of insurance. The lack of identifying and verifying highly predictive rating variables for auto insurance pricing to help reduce the need for rate subsidization is the general problem to be focused on with the study. A quantitative correlational analysis was conducted to address the research questions and determine the predictive strength of key independent rating variables. The next section of the research study will discuss the role of the

researcher, the dataset with the appended third-party information, the population and sampling, as well as the analytical approaches and the verification of reliability and validity of the research.

Section 2: The Project

The quantitative study is focused on determining the criminal background variable predictive strength in estimating loss costs for auto insurance ratemaking. Based on the information from the auto property and casualty literature review, there is a gap in the research for how risky personal behavior may translate to risky driving behavior. The data and analysis presented in The Project section will provide the details for the analytical approach and verification process for modeling the criminal driver background variable impact with an auto insurance book of business.

The Project section encompasses the full range of procedures necessary to conduct the research study, the overview of the purpose of the study, the role of the researcher, driver data access, research methodology and design, data population and sampling, data collection, analytical approaches, and a review of the validity and reliability of the data. This section is essential to fulfill the requirements of a properly executed quantitative research study.

Purpose Statement

The purpose of this quantitative study is to examine if an individual's criminal background is associated to a person's future driving behavior and if it is predictive of future loss costs. Risk segmentation attempts to mitigate pricing subsidization between lower and higher risk drivers, which supports market efficiency and addresses the increase of social risk cost and the loss of equity (Duan et al., 2018). The research by Hoy (1982) showed higher-risk classes receive coverage at an actuarially uniform premium, while lower-risk classes receive less than full coverage at an actuarially uniform premium. Hoy (1982) highlighted if the proportion of lower-risk class falls below a determined breakeven threshold, the competitive equilibrium will no longer be a no-subsidy segmentation approach. Lower-risk classes would subsidize the

higher-risk classes, which would not meet management and shareholder expectations in the current economic and competitive climate.

Private passenger auto premiums serve two primary purposes for a property and casualty carrier: one being the premium should be able to cover the expense and risk obligation, and the other should price the insurance premiums equitably for specific risk classes (Cova et al., 2016). Individual criminal background information has only recently become available as third-party data available in public and private databases. Including individual criminal background as a rating or underwriting variable would potentially allow carriers to further segment higher-risk drivers for accurate pricing, which would allow carriers to balance profitability with equitable pricing in the market.

Role of the Researcher

The researcher's quantitative study role is to work with the insurance carrier's data science team and a third-party vendor to provide validated data to append to an existing auto insurance dataset for statistical analysis. The study's selected approach is a correlational design to support the potential impact of individual criminal background behavior on auto insurance loss costs. With the proposed dataset and the amount of data available, a quantitative research method is the preferred analytical approach for statistical analysis (Creswell & Creswell, 2018).

To restrict personal identifiable information (PII) data from being used in the analyses, the carrier's data science team conducted the data extract from the existing non-standard auto data warehouses. The dataset from 2014 to 2017 included 20 states with individual departments of correction contributing to the third-party vendor database. Data from 2014 – 2017 was used to ensure developed incurred losses and earned exposures would be available for the loss cost calculations and determine if the criminal background variable has predictive strength in

estimating pure premium**.** The data science team appended criminal background variables to 448,755 non-standard auto insurance records, which also contain an insurance-based credit variable so the researcher could determine if there is a correlation between the two driver rating variables. The criminal background variable will also be introduced to the current Generalized Linear Model both with and without the insurance-based credit variable to determine the model's predictive strength when including the additional variable. Based on the correlational design, outside of appending the criminal background data to the individual driver records, no additional data elements or PII data are included with the finalized records.

Multiple linear regression using Generalized Linear Models will be the modeling approach using the Pearson correlation coefficient *r* for the parametric statistic (Krzywinski & Altman, 2015). The Generalized Linear Model is the current actuarially approved method for pricing the auto line of business filed with the state departments of insurance. Both independent variables, insurance-based credit score and criminal background, are evaluated for contributions to predicting the dependent variable, expressed as loss costs, and in determining correlation. An evaluation of each research question and the supporting hypotheses is necessary to determine if statistical significance exists.

Participants

The research study dataset was extracted from an existing property and casualty auto insurance pool of non-standard auto policyholders. No human participants or their associated PII data were made available during the study. Three years of experience data were selected to allow for earned exposure and incurred loss development for the pure premium calculation. Witt (1974) stated full credibility implies the current loss experience was adequate to conducting ratemaking based on the sample of experience alone. If it were possible to study a more

considerable amount of similar data, the rates developed would be no more reliable than the rates based on the selected loss data given 100 percent credibility.

An extract from the insurance record data included 20 states, and all records with complete driver information are in the data extract. The data science team appended the most recent criminal conviction occurrence for all drivers with an event. The third-party database contained data from all 50 states and also included federal criminal convictions. The third-party criminal history is independent of driving history captured by the state departments of motor vehicles, and the conviction data are cross-referenced and validated to ensure no overlap in the external data sources.

Research Method and Design

The complexity of current property and casualty business models requires research designs, which provide experimental control using large amounts of data from an organized set of policy records and quantitative analysis (Collins et al., 2009). Design methods are foundational studies to assess and develop approaches based on historical research (Lock & Seele, 2015). Collins et al. (2009) stated the research design should advance the analytical methodologies and support the underlying hypotheses and provide potential improvements. A quantitative research approach was selected over a qualitative research approach because the archival dataset has historical numerical and categorical data instead of records based on surveys and observations (McCusker & Gunaydin, 2015).

Discussion of Method

Brockett and Golden (2007) stated correlations are a statistical means of determining the relationship between a risk factor and the likelihood of a loss event. A quantitative correlational research approach is applicable based on the theoretical framework, which will test the

relationship between the dependent variable, pure premium, and the independent variables, criminal background, and insurance-based credit score. The statistical analysis will determine the relationship between the variables and the strength and contribution of each variable to the model (Brannon, 2016).

The variables and the research questions' desired outcomes determined the quantitative approach over a qualitative or mixed methods approach. The population of archival numeric and categorical data is from the carrier's policy data warehouses. The data science team also appended claims and third-party data for modeling. In contrast, qualitative approaches for research studies provide a method to describe and document analytical overviews from small study samples to extrapolate to the broader general population (Siedlecki, 2020).

Discussion of Design

A correlation design will help establish if a relationship is present between criminal background, insurance-based credit scores, and pure premium. The research study defines the nature and the magnitude of the relationship between the variables (Salkind & Frey, 2020). Bosco et al. (2015) also underscored the facts citing a quantitative correlational approach supports research involving both financial services and pricing studies as the method is trying to find a single answer to the proposed hypothesis. The experimental and descriptive quantitative approaches were not a good fit for the proposed research study. The approaches are directed toward research subjects in natural environments and could be considered more preliminary research for simple statistical designs (Edmonds & Kennedy, 2017).

Salkind and Frey (2020) stated the experimental and quasi-experimental approaches to quantitative research require independent, dependent, and control variables, including a control procedure to distinguishing a cause and effect relationship. Edmonds and Kennedy (2017)

outlined the descriptive design as more in line with a survey directed to a random sample within a target population. Researchers conduct surveys to observed trends and behaviors within the sample population, and the proposed research study is to determine the correlation between variables (Bosco et al., 2015). The experimental, quasi-experimental, and descriptive designs are not compatible with the dataset and the hypothesis developed for the correlational study. The experimental, quasi-experimental, and descriptive designs would also require manipulating the data collection (Siedlecki, 2020).

Key Independent and Dependent Variables. The dataset represents rating variables, including the most recent appended criminal conviction events to the insurance records from a non-standard auto carrier. The segmentation of criminal events into 23 categories for misdemeanor and felony convictions are determined by the vendor and standardized across the reporting events from the individual departments of correction. There was an eight percent hit rate for drivers with criminal background data from the population of the 20 states selected for the study. Nagrecha et al. (2018) stated given the limitations of using the probability of criminal activity alone for a policy, the research requires metrics, which measure the overall impact on loss costs and should include all underwritten policies with the associated losses and earned exposures. A significant component of the study is all drivers who did have a criminal conviction in their background would be offered insurance coverage at the proper premium and were not underwritten away as an unacceptable risk (Kurlychek et al., 2006).

The insurance-based credit score is a well-established rating variable and is ranked within the top three most important loss predicting variables when evaluating individual line coverages during multivariate modeling (Brockett & Golden, 2007). The direct statistical relationship between the primary named insured's financial stability and future loss events can help the

carrier determine underwriting rules and rate segmentation (Stiff et al., 2019). Rating categories determine buckets of insurance-based credit scores, and the study used a vendor-filed and approved model. There were 14 credit categories defined by the vendor for analysis and rating purposes, with no individual scores being present in the dataset and only the credit bucket associated with the insurance record available for the study.

For this correlational design, the modeling of two independent variables against the target variable, or the dependent variable, determines the relationships between the variables within the non-standard auto insurance policy population (Curtis et al., 2016). The pure premium was calculated for each insurance record by summing the incurred losses and dividing by the sum of the earned exposures, divided by 365 days to adjust for daily earnings. This method allows the models to determine the projected values for insurance loss costs (Werner & Modlin, 2016). Using the pure premium method allows for a frequency and severity approach using Generalized Linear modeling (Frees et al., 2014; Garrido et al., 2016; Jeong et al., 2017).

Summary of Research Method and Design

A correlational analysis based on a quantitative approach was selected because of the historical numerical data to determine the predictive strength of the primary independent variable with the dependent target variable. A quantitative design was determined to be more appropriate than a qualitative, observational approach to collecting data. For successfully implementing a new variable into the underwriting or ratemaking solution, actuarially sound statistical support needs to accompany any filing made with the state departments of insurance (Spedicato et al., 2018).

Population and Sampling

Lock et al. (2016) stated all individuals or subjects of interest, including all of their associated elements, are defined as the population. For financial and pricing studies, a researcher may be working with a sample or subset from a large population of available recorded occurrences. The dataset and records selected for modeling determine the difference between a sample and the full population (Lock & Seele, 2015). The quantitative study's target modeling population is a subset of insurance records between 2014 and 2017, from 20 states with fully developed losses and earned exposures required to determine a more accurate forecast of the dependent variable, pure premium.

Discussion of Population

The researcher had a large population of non-standard auto records from a property and casualty insurance carrier to append the criminal background variable for modeling. Auto insurance carriers retain insurance records for six years to meet state regulatory requirements before archiving the data for additional research purposes. Data Science and actuarial teams prefer to use eight to ten years of data brought to current rate level for predictive modeling. The additional data does not add incremental value to smaller-scale studies outside of ratemaking (Quan & Valdez, 2018).

Discussion of Sampling

The convenience sample of insurance records included all policies for the 3 years between 2014 and 2017 from 20 states (Edmonds & Kennedy, 2017). Using the most recent three years of experience data allowed for loss and exposure development and facilitated the most recent criminal conviction event to be appended to the dataset. Etikan et al. (2016) stated in cases where the data were readily available, and convenience sampling is a suitable approach as

the records from the target population are homogeneous. There would be no difference from a random sample. Included in the three years of data are 448,775 insurance records with rating variables for modeling. The team ordered criminal history reports for each driver on the policy. The data appended included the single most recent criminal conviction on record or a clean criminal history report for each driver.

Summary of Population and Sampling

The full population of ten years of non-standard auto insurance records provided a readily accessible pool of risks for appending the criminal background variable to driver records. The researcher and the supporting data science team had the required experience grounded in insurance-based pricing and modeling knowledge to select an appropriate and credible dataset using a convenience sampling approach for the quantitative study (van Rijnsoever, 2017). During the three years with developed losses and exposures, all insurance data are available for modeling with a criminal background report for each listed driver. The characteristics of the convenience sample used for modeling were homogeneous as the elements for rating and evaluating each insurance record are the same (Etikan et al., 2016).

Data Collection

Cotteleer and Wan (2016) stated data acquisition phases could be shorter by using corporate archival information as a reliable way to bridge the gap into developmental research with large amounts of standardized data available for studies. The researcher and the carrier's data science team had the required industry expertise to select an appropriate and credible dataset, normalized, and summarized at the risk record level. The use of corporate archival data allowed for the losses and exposures to develop to provide a more accurate dependent variable, pure premium, to support statistical analysis.

Instruments

For the non-standard auto dataset, the carrier's data science group will work with the researcher to determine the necessary elements for the data extract and the appropriate period. The data science group will query the necessary variables so the researcher will not have access to PII data at the policy level. The team will provide the data to the researcher in both a Statistical Analysis System (SAS) file, and a Comma Separated Values (CSV) file. SAS OnDemand for Academics is available for Generalized Linear Modeling (GLM), and IBM Statistical Package for Social Sciences (SPSS) is available for variable categorization and the initial exploratory data analysis. The data variable library is easily accessible from a Microsoft Excel file. Surveys and interviews are not needed to complete this study.

Data Collection Techniques

The data science analyst will write a SAS query to pull the necessary variables from the carrier's Policy Enterprise Data Warehouse (EDW), Billing Data Warehouse, and Claims Data Warehouse. The analyst will also consolidate the line coverage level variables up to the policy level. The raw data stored in the various EDWs was normalized for insurance carrier actuarial pricing and reserving and at the lowest level possible for modeling and the policy number replaced with a record identification number.

The team will append criminal history events for each driver in the population, and the activity codes are aggregated at the record level. The criminal background reports are classified as Federal Credit Reporting Act (FCRA) information, and all report orders have the consumer's approval at the new business quote. Activity codes were dividing into logical categories for future analysis at the event level. No-hit reports warranted a separate category for no criminal background activity. The approach for ordering reports on all drivers follows the same

guidelines, which are in place for ordering an insurance-based credit report on all risks being quoted for insurance so ensure fair treatment of individual risks.

Data Organization Techniques

Three years of archive data from 2014 to 2017 were available in both CSV and SAS file format for analysis purposes. The CSV file is formatted to import into SPSS to categorize nominal variables and ensure all data fields are ready for modeling. New variables were created in SAS to aggregate incurred losses and earned exposures at the record level. Earned exposures needed to be divided by 365 (days) to calculate the pure premium variable by dividing total incurred losses by earned exposures for each record.

The data file storage includes a secure personal hard drive, a flash drive, and the SAS file in SAS OnDemand cloud storage. After the research publication, the destruction of all data files is necessary to remain in compliance with the Non-Disclosure Agreement with the insurance carrier. The data science team has the original raw dataset and can retain the information to comply with federal guidelines. The overview of the approach and the data elements used will allow for the replication of the analysis.

Summary of Data Collection

Corporations with growing archival data provide new analytical opportunities, which researchers need to leverage to gain significant product development insights and enhanced business cases. While some academics are more inclined to conduct organic field studies, more researchers are finding having access to increased amounts of new data allows for advanced business analytics with a reduced level of risk and shorter timelines (Cotteleer & Wan, 2016). The three years of archival data extracted by the carrier's data science group and the appended driver background conviction events will allow for multiple linear regression analysis of the

target variables with pure premium. Individual surveys or interviews are not needed to develop the dataset, and the extract did not include any PII data. The archival dataset will be the basis of the statistical analysis for each research question, and the destruction of all data files will occur upon publication of the study. The carrier's data science team will retain a copy of the dataset to comply with federal guidelines.

Data Analysis

Each variable will require exploratory data analysis to understand the data structure and verify assumptions before conducting inferential statistics, which will drive the types of statistical tests and the appropriate models. The dependent variable for insurance modeling does not have a normal distribution because each policy record will not have a loss event to calculate individual loss costs. Historically, claims frequency data follows a Poisson or negative binomial distribution (Yip & Yau, 2005). For skewed independent variables, as long as the residuals have a normal distribution, those variables can be used in linear regression models.

Variables Used in the Study

The dataset is archival data from an insurance carrier, and there are no variables with missing data. During exploratory data analysis (EDA), an outlier review will determine if any data needs to be imputed or truncated. The criminal background variable is a categorical variable, which follows a similar zero-point, right-skewed distribution as the dependent variable. The insurance-based credit score variable converts to a categorical grouping as the carrier files a credit group instead of an actual score, which follows a normal distribution. The incurred losses divided by the earned exposures create the pure premium variable for each risk. The results skew-right due to the large number of zero-incurred claims. A log transformation converts the pure premium to a normal distribution to meet the assumptions of inferential statistics (Bamattre

et al., 2017). For categorical variables, Cramer's *V* is a nonparametric measure to test the strength of the relationship between two variables and is appropriate for testing the criminal background and insurance-based credit score variables (Morgan et al., 2013).

The correlational study includes modeling from a one-parameter exponential family of distributions, which can accommodate zero-point distributions (Haberman & Renshaw, 1996). The logistic regression model, where the dependent variable is dichotomous, is the consistent choice for auto insurance modeling (Klieštik et al., 2015). Since most of the study's available variables are categorical, using binned-variables and a Poisson Regression Generalized Linear Model approach is an appropriate method for hypothesis testing insurance claims event data (Goldburd et al., 2020; Little, 2013). Additional variables for the Generalized Linear Model portion of the study include age, past driving violations, financial responsibility indicators, and months of prior insurance. Interaction variables, which are the product of multiplying two independent variables, are also available for modeling the strength of the two independent variables (Allison, 1977).

Table 1

Correlational Model Variables

Correlation Analysis Model Variables			
Variable	**Description**	**Variable Type**	**Data Type**
Criminal Background	Individual misdemeanor and felony convictions	Independent Variable$_1$	Nominal
Insurance-Based Credit Score	Individual credit score developed to predict future losses	Independent Variable$_2$	Nominal
Pure Premium	Incurred Losses divided by Earned Exposures	Dependent Variable	Scale
Insurance Variables for Generalized Linear Model			
Driver Minimum Age / Violations Interaction	Interaction variable: Youngest Driver and their violations	Independent Variable	Scale
Driver Minimum Age / Years Licensed Interaction	Interaction variable: Youngest Driver and their years licensed	Independent Variable	Nominal
Household Composition	Matrix variable: Number of Vehicles, Number of Drivers, Age of Youngest Driver	Independent Variable	Nominal
Insured Age/Insurance-Based Credit Score Interaction	Interaction variable: Age and Insurance-Based Credit Score	Independent Variable	Nominal
Non-Sufficient Funds / Late Pay Interaction	Interaction variable: Number of Unfunded Payments and Number of Late Payments	Independent Variable	Nominal
Prior Months Insurance	Number of Months of Prior Insurance Coverage	Independent Variable	Nominal
Recent At-Fault Accident	Time between Effective Date of the Policy and the Most Recent Accident	Independent Variable	Nominal
Recent Minor Violation	Time between Effective Date of the Policy and the Most Recent Violation	Independent Variable	Nominal
State / Coverage Interaction	Interaction Variable: State and Line Coverage	Independent Variable	Nominal
SR-22 Filing	Presence of a Financial Responsibility Filing	Independent Variable	Nominal
Territory / Line Coverage Group	Matrix Variable: Zip Code of the Policy Record and Coverage Purchased	Mitigating Variable	Nominal
Vehicle History	Event History of the Vehicles on the Policy Record	Independent Variable	Nominal

Hypothesis 1

The lognormal transformed dependent variable, pure premium, comes close to a normal distribution and compares the count events with a Poisson Regression Generalized Linear Model approach (Goldburd et al., 2020). The model test uses an alpha set to the 0.05 level to determine an association between the criminal background variable and the dependent variable. The null hypothesis, stating there is no incremental predictive strength for criminal background, will be rejected if $p < 0.05$, which is statistically significant (Leech et al., 2015).

Hypothesis 2

The dataset has a large population sample size for a Chi-Square and Cramer's V to test the relationship between criminal background and insurance-based credit score. For the Chi-Square Test of independence, the significance level is set to $p < 0.05$ to assess whether to accept or reject the null hypothesis and determine if the two independent variables of interest are unrelated (Morgan et al., 2013). The Cramer's V statistic provides information concerning the strength of the relationship between criminal background and insurance-based credit score and is similar to reading a correlation score with an outcome close to zero indicating no relationship (Weihs et al., 2018).

For the second set of hypotheses statements, the initial iteration of the Poisson Regression Generalized Linear Model will be run with the criminal background variable and tested at a significance level of $p < 0.05$. The same version of the model will run with the insurance-based credit score, a significance level of $p < 0.05$, and the criminal background variable removed. Variable selection is an essential step in statistical model evaluation, providing the simplest model for the data set provided to improve prediction outcomes and retaining the goodness of fit.

The Akaike's information criteria (AIC) can evaluate each iteration of the model and select the model with the smallest AIC value (Kimura & Waki, 2018).

Summary of Data Analysis

Insurance rating variables are distinctive due to the event-specific, right-skewed data, present with both the dependent variable and the contributing independent variables used for modeling. The ability to use inferential statistics and models aligned with categorical variables helps provide the required analytical support used in ratemaking (Ranganathan et al., 2017). The statistical review of the criminal background variable must support the purpose statement and needs to determine the magnitude and strength of the relationship to pure premium. The analysis also needs to address if an interaction exists with the most predictive independent rating variable, the insurance-based credit score.

The research questions and the associated hypotheses framed by the Chi-Square and Cramer's V tests, test the level and strength of the relationship between two variables, are appropriate for testing the criminal background and insurance-based credit score variables with pure premium (Morgan et al., 2013). The Poisson Regression Generalized Linear Model uses the lognormal transformed dependent variable, pure premium, to compare the count events (Goldburd et al., 2020). Evaluating and ranking the models with a combination of the smallest AIC value and fewest number of variables from the data set supports the variable selection process to improve overall prediction performance (Kimura & Waki, 2018).

Reliability and Validity

The study's causal-comparative design attempts to find specific relationships between the independent variable, criminal background, and the dependent variable, pure premium. Reliability and validity are review processes, which attempt to evaluate the overall quality of the

data collection approach and the supporting measurements and statistical analyses (Heale & Twycross, 2015). Reliability addresses the consistency of the study's measurements, and validity indicates how accurately the method addresses the measurement approaches (Uraschi et al., 2015).

Reliability

The study's design addresses reliability during the data collection and data analysis phases of the research (Yin, 2018). The carrier's data science will query the archival dataset in the study consistent with an auto insurance actuarial approach. An enterprise data warehouse team cleanses and normalizes the data before storage. Accuracy and consistency are necessary to ensure repeatable tests at different times and for different sample populations (Barnes et al., 2018). Even though the archival dataset is not public, insurance carriers are held to the same standards when filing actuarial exhibits with the individual departments of insurance. The actuarial support exhibits follow a similar format across carriers.

The study introduces a second independent variable, insurance-based credit scores, to explore possible alternative explanations during statistical testing (Yin, 2018). Comparing each variable within the Generalized Linear Model and against each other may help explain the why or the how of the potential outcomes (Uraschi et al., 2015). The archival data for the study allows for more advanced statistical modeling with a larger dataset and quickly appending available third-party data for testing purposes. The dataset used in micro-organization research can increase the statistical significance by reproducing the study with additional information (Barnes et al., 2018).

Validity

The validity determines the research methodology's accuracy and whether the approach produces the intended results (Watson, 2015). Heale and Twycross (2015) stated criterion validity explains the correlation between the variables of interest, the statistical testing approach, and whether the testing results correlate with previous test results. Both insurance-based credit score and criminal background reviews for correlation with the dependent variable pure premium. The insurance-based credit score is currently the most predictive variable in auto insurance pricing and serves as the baseline for a test with pure premium (Bärtl & Krummaker, 2020; Brockett & Golden, 2007; Golden et al., 2016; Insurance Information Institute, 2019; Lee et al., 2005; Morris et al., 2017). An archive dataset, the selected research variables, and a Generalized Linear Modeling approach will determine a correlation between the variables.

Content validity determines if the selected instrument for measurement used for a study is appropriate for the data and the research design (Creswell & Creswell, 2018). The concerns involving content validity are not statistical and instead are focused on the amount and level the archive data used in the study is enough to determine sample behavior (Watson, 2015). The insurance loss cost study uses a historical dataset instead of an instrument to measure the criminal background behavior.

The construct validity involves the precision of the forecasts and measurements for the outcomes of the research questions posed (Heale & Twycross, 2015). The most significant challenge for construct validity is matching the categorical variables with the correct statistical tests. The data science team and recent publications on insurance modeling offered guidance for appropriate approaches. Bian et al. (2018) supported using a Generalized Linear Model for the validation process of testing insured's loss costs and matched criminal background behavior. For

categorical or binned data, Poisson Regression Generalized Linear models are appropriate and mitigate construct validity concerns (Little, 2013).

Summary of Reliability and Validity

The study design and quantitative approach with an archival dataset support the research questions and apply them to the current insurance industry pricing methodology. There is an expectation the archive data will help moderate the risks associated with reliability and validity, and the study is using the same statistical models from research published in academic journals to evaluate rating variables (Cather, 2018; Frees et al., 2014; Garrido et al., 2016). The data science team will also provide suggestions for ensuring the study approach is accurate.

Reliability and validity reviews are critical steps, and the study should strive to accomplish to ensure the research is replicable (Yin, 2018). Reliability specifies the level and degree the study measurements are consistent, and the research is replicable. Validity tries to address the accuracy of the modeling and the actual outcomes of the study. The use of archival data may decrease the overall risks associated with reliability and validity and may improve the quality of the study's outcomes.

Transition and Summary of Section 2

The quantitative, correlation study examines the potential relationships between the independent variables, criminal background, and insurance-based credit score, with the dependent variable pure premium. The purpose of the study, research questions, and the hypotheses are relevant to the research approach and provide the analysis direction. The Generalized Linear Modeling approach supports the research questions and hypotheses and is similar to other published auto insurance studies referencing rating variables (Tan et al., 2015;

Tan, 2016a, 2016b). The data science team mitigates concerns involving reliability and validity with the use of archive data.

The final section addresses the outcomes of running the analyses for the research questions, and the results of measurements and modeling assessments. A brief study overview, an examination of the research questions, and the related hypotheses and the statistical significance of the outcomes occur in the final section of the study. The section will outline the auto insurance industry's applications, the recommendations for immediate and future actions, and additional research to pursue. The section will conclude with an evaluation and reflection of the study and the overall research process.

Section 3: Application to Professional Practice and Implications for Change

The research was completed to determine if there are potential rating variables, which will improve the predictive models used for auto insurance pricing to decrease the need for rate subsidization. Carriers have access to more robust data, providing advanced business insights for more accurate model outcomes. Including the criminal background variable to the existing Generalized Linear Model allowed for the testing of relationships between the independent variables and the overall goodness of fit of the various model iterations (Kimura & Waki, 2018).

The dataset with non-standard auto insurance records was used to test the research questions and the supporting hypotheses. The section will include an overview of the research approach and how it applies to the finance field. The analysis outcomes provide a detailed discussion of the descriptive statistics, tests conducted, and the hypotheses' results for the research questions. The Akaike Information Criteria, measuring the goodness of fit with in-sample prediction error for the Poisson Regression General Linear Models, had the lowest AIC score when both of the key independent variables were included in the model.

Overview of the Study

The purpose of the quantitative, correlational study was to test the predictive relationship of criminal background with insurance-based credit scores and pure premium to determine if pricing segmentation opportunities exist. Research on the industry reveals carriers using average pricing are under pressure to find risk segmentation opportunities to provide a better match for an individual's risk profile based on rate differentiation (Lass et al., 2016). Investing in premium development design based on updated risk factors allows for heightened competition and decreased rate subsidization.

The study expanded on business process innovation research and addressed the criminal background variable's verification and rating use (Ohlsson et al., 2015). The first research question addressed the criminal background variable's predictive impact on insurance loss costs, which were modeled with the dependent variable, pure premium. Criminal background was included in the Generalized Linear Model to determine if the variable improved the goodness of fit for predicting future pure premiums. The pure premium model had a better overall fit when the criminal background variable was included with other highly predictive variables already used for premium development.

The insurance-based credit score was also compared to criminal background to determine if there was an association between the variables. For future department of insurance filings, the correlation between a person's criminal background and credit score would be a question from a regulator. While the anticipated correlation was present, the association's overall strength was small and weak, allowing both variables to be used in modeling.

The last set of analyses was a series of Poisson Generalized Linear Models run with the dataset's strongest predictor variables to determine the new variable's potential influence for forecasting pure premium. Insurance-based credit score has a slightly higher correlation to pure premium, while the criminal background variable within the full model had a better goodness of fit response than credit when both variables were tested alone.

Presentation of the Findings

The study used a sample of archival 448,755 insurance records from 2014 to 2017. The single most recent, highest ranked, criminal background event was appended to the records to determine if the predictive model fit improved with the additional data. A random sample was run to split the dataset into sets for training and testing the models. Table 2 displays the 75/25

split, with 335,477 records used for training, 112,275 used for testing, with no excluded records. There was one identification variable for each model, either insured loss or pure premium, along with iterations of the eleven categorical variables.

Table 2

Auto Insurance Training and Test Dataset

Sample	N	Percent
Training	336,477	75.0%
Testing	112,275	25.0%
Valid	448,775	100.0%
Excluded	0	0.0%

Table 3 displays the descriptive statistics for the variables in the dataset. The Data Science team provided a scrubbed dataset and all 448,755 records have valid data in all fields with no missing or miscategorized data so no imputation was necessary. The Pure Premium variable was a skewed-right, scale variable, and was converted with a natural logarithm transformation for analysis purposes.

Table 3

Descriptive Statistics for Insurance Carrier Variables

Descriptive Statistics for Insurance Carrier Variables

	Criminal Background	Insurance-Based Credit Score	Pure Premium	Household Composition	Insured Age/Insurance-Based Credit Score Interaction	Non-Sufficient Funds / Late Pay Interaction	Prior Months Insurance	Recent At-Fault Accident	Recent Minor Violation	SR-22 Filing	Territory	Vehicle Hisory
N = Valid	448,755	448,755	448,755	448,755	448,755	448,755	448,755	448,755	448,755	448,755	448,755	448,755
N = Missing	0	0	0	0	0	0	0	0	0	0	0	0
Mean	2.04	9.30	781.76	5.92	2.79	1.91	2.22	2.78	2.86	1.16	12.67	1.03
Median	0.00	10.00	0.00	2.00	3.00	2.00	2.00	3.00	3.00	1.00	12.00	1.00
Skewness	8.429	-0.394	5.035	0.987	-2.534	0.982	0.377	-2.340	-3.133	1.872	0.152	5.433
Std. Error of Skewness	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
Kurtosis	70.418	-0.936	25.533	-0.306	4.921	0.613	-1.533	4.331	9.369	1.505	-1.264	27.518
Std. Error of Kurtosis	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
Minimum	0	1	0	1	1	1	1	1	1	1	1	1
Maximum	97	14	26096	18	3	5	4	3	3	2	27	2

Tables 4 and 5 explored the frequency distributions for the two independent variables of interest. In Table 4, the criminal background events were separated into groups of events based on an individual having a clean record, felony, misdemeanor, or one of the fraud events. Within felony

and misdemeanor groups, the events were further segmented into categories increasing the granularity of the data being analyzed. The frequency distribution was heavily skewed-right due to the high number of clean records in the dataset.

Table 4

Criminal Background Frequency Distribution

Criminal Background Frequency Distribution

Group	*Code*	*Description*	*Frequency*	*Percent*	*Cumulative Percent*
No Hit	0	Clean Record	392,377	87.4%	87.4%
Felony	01 (1)	Public Peace & Welfare	1,411	0.3%	87.8%
	02 (1)	Government, Law Enforcement & Justice	587	0.1%	87.9%
	03 (1)	Drugs	6,413	1.4%	89.3%
	04 (1)	Property	4,049	0.9%	90.2%
	04 (2)	Person	3,021	0.7%	90.9%
	94 (1)	Unspecified	569	0.1%	91.0%
Misdemeanor	01 (2)	Public Peace & Welfare	769	0.2%	91.2%
	02 (2)	Government, Law Enforcement & Justice	468	0.1%	91.3%
	03 (2)	Drugs	741	0.2%	91.5%
	04 (2)	Property	1,957	0.4%	91.9%
	07 (2)	Person	11,824	2.6%	94.5%
	94 (2)	Unspecified	1,828	0.4%	94.9%
Unacceptable	05 (1)	Financial Fraud (Felony)	8,505	1.9%	96.8%
	05 (2)	Financial Fraud (Misdemeanor)	6,465	1.4%	98.3%
	06 (1)	Insurance Fraud (Felony)	4,410	1.0%	99.3%
	06 (2)	Insurance Fraud (Misdemeanor)	3,362	0.7%	100.0%
Total			448,755	100.0%	

Table 5 displays the insurance-based credit score distribution, which was separated into credit score groups. The insurance-based credit scores were determined by a proprietary model provided by a vendor, and the consumer must provide permission for the carrier to run the score for rating and underwriting purposes.

An insurance-based credit score is similar to a consumer credit score, but the insurance score range is much wider from top to bottom. The insurance-based credit score was developed to determine if a consumer will incur a loss and file a claim. For the insurance dataset, each state determines how No Hit and No Score reports will be handled for rating as consumer treatment

must be consistent. Those scores will be mapped to an average Score Group providing a more normal distribution for the dataset.

Table 5

Insurance-Based Credit Scores Frequency Distribution

Insurance-Based Credit Score Frequency Distribution

Score Group	*Frequency*	*Percent*	*Cumulative Percent*
0-499	6,182	1.4%	1.4%
500-524	11,373	2.5%	3.9%
525-550	22,797	5.1%	9.0%
551-568	29,247	6.5%	15.5%
569-585	20,203	4.5%	20.0%
586-605	25,918	5.8%	25.8%
606-626	25,559	5.7%	31.4%
627-649	27,979	6.2%	37.7%
650-676	34,093	7.6%	45.3%
677-719	48,738	10.9%	56.1%
720-818	65,406	14.6%	70.7%
819-997	16,927	3.8%	74.5%
No Hit	17,694	3.9%	78.4%
No Score	96,639	21.5%	100.0%
Total	448,755	100.0%	

The research questions examined the criminal background variable's predictive strength with the existing model variables to project pure premiums. Criminal background was also reviewed against the insurance-based credit score variable to determine if there was a correlation between the two variables and the strength of the association. The final set of analyses introduced the criminal background variable to the Poisson Regression Generalized Linear Model to determine if including the additional variable improved the model's fit. Improving the model fit would decrease the reliance on price subsidizations by charging the right premium for the underlying risk.

The results for each research question and related hypothesis were addressed and supported by the statistical test results. The analytical outcomes will note any data outliers or discrepancies, and any potential impact on the approach. The association test results in research question one were statistically significant, and the null hypothesis could be rejected. The nonparametric correlational test for research question two between criminal background and insurance-based credit score was also statistically significant, and the null hypothesis could not be rejected, although the association was small and weak. There was an improved goodness of fit for the model for question three when introducing the criminal background variable and removing the insurance-based credit score variable, and the null hypothesis could be rejected. The findings are consistent with other supporting studies, which suggest carriers review the increase in precision and complexity of adding new variables to the pricing model before introducing the model to the market (Lass et al., 2016).

Hypotheses 1

The first research question introduced the criminal background variable to a variation of the current Generalized Linear Model using the strongest predictive variables to determine if increased model fit was possible.

RQ1. What is the predictive impact of an individual's criminal background on auto insurance loss costs?

$H1_o$ = There is no additional incremental predictive ability of the current pure premium (DV) model with the addition of the criminal background data.

$H1_1$ = There is additional incremental predictive ability of the current pure premium (DV) model with the addition of the criminal background data.

A Poisson Log Linear Generalized Linear Model was run on the testing dataset with the identified baseline predictor variables outlined in Table 1 with both the insurance-based credit score and the criminal background variables. Table 6 displays four models with the Akaike Information Criteria (AIC) score, which estimates the model's relative quality given a set of data and the number of predictor variables used to create the model. All of the models have statistical significance with $p < 0.001$. Model 4, With Insurance-Based Credit Score and Criminal Background, had the lowest AIC value indicating a better fit than Model 2, which included only the Insurance-Based Credit Score (Kimura & Waki, 2018).

Table 6

Relative Quality of Statistical Models

Model	Method	AIC[a]	Log Likelihood	Likelihood Chi Square[b]	Significance	Predictor Variables
1. Model without Insurance-Based Credit Score	Poisson Log Linear	205598.878	-102741.439	5287.838	0.000	9
2. Model with Insurance-Based Credit Score	Poisson Log Linear	205159.662	-102500.831	5769.054	0.000	11
3. Model with Criminal Background and without Insurance-Based Credit Score	Poisson Log Linear	204710.470	-105588.238	6194.247	0.000	10
4. Model With Insurance-Based Credit Score and Criminal Background	**Poisson Log Linear**	**204287.414**	**-102055.707**	**6659.303**	**0.000**	**12**

a. Information criteria in smaller-is-better presentation

b. Compares fitted model with intercept-only model

The AIC value, 204287.414, was the lowest for Model 4, and also had a $p < 0.001$, which allowed the null hypothesis to be rejected as there was an incremental improvement in the model fit and predictive ability with the addition of the criminal background variable. Model 4, with the corresponding variable coefficients, is displayed in Appendix A.

In Logistic Regression, the Wald Chi-Square test can be used to assess whether or not a variable is statistically significant within the selected model. The output for Model 4: Poisson Log Linear Model with Insurance-Based and Criminal Background had the Wald Chi-Square statistics generated for the categorical variables in the testing or hold-out data set. Table 7 displays the predictive variables based on higher Wald Chi-Square values and lower p -values. The criminal background variable had the highest Wald Chi-Square value, Wald X^2 (9, N =

112,275) = 951.206, $p < 0.001$. For the testing data set, the insurance-based credit score variable ranked fourth in predictive contributions with a Wald X^2 (13, N = 112,275) = 387.964, $p < 0.001$.

Table 7

Wald Chi-Square Analysis of Insurance-Based Credit Score and Criminal Background with Incurred Losses

Wald Chi-Square Analysis of Criminal Background and Insurance-Based Credit Score with Incurred Losses

	Wald Chi-Square	*df*	*Significance*
(Intercept)	885.569	1	0.000
Criminal Background	951.206	9	0.000
Household Composition	893.423	16	0.000
Recent At-Fault Accident	468.274	2	0.000
Insurance-Based Credit Score	387.964	13	0.000
Territory / Line Coverage Group	266.844	26	0.000
Prior Months Insurance	262.694	3	0.000
Recent Minor Violation	77.359	2	0.000
Non-Sufficient Funds / Late Pay Interaction	57.650	4	0.000
Vehicle History	54.033	1	0.000
SR-22 Filing	34.246	1	0.000
Insured Age/Insurance-Based Credit Score Interaction	27.795	1	0.001
Driver Minimum Age / Violations Interaction	25.819	4	0.012

Dependent Variable: Incurred Loss

While all of the variables selected for modeling were statistically significant, the question of model complexity was also reviewed. The number of variables increased by more than one when including insurance-based credit score as the variable is also used in an interaction term. Using a combination of AIC value, Wald Chi-Square, and the number of predictor variables helps develop a better-balanced model selection for expected outcomes.

Hypotheses 2

The second research question addressed the concern there may be a significant correlation between the most predictive rating variable, insurance-based credit score, and criminal background.

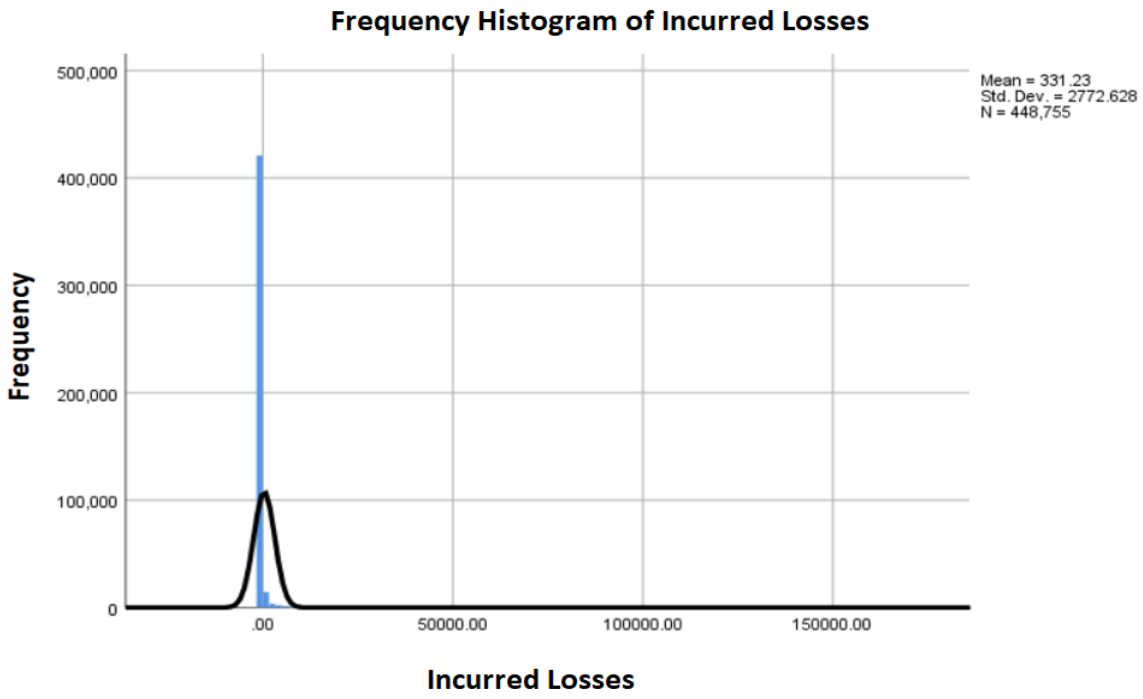
RQ2. What is the relationship between the insurance-based credit score and criminal background?

$H2_o$ = There is no statistically significant correlation between an individual's criminal background and insurance-based score.

$H2_1$ = There is a statistically significant correlation between an individual's criminal background and insurance-based score.

Figure 2

Frequency Histogram of Incurred Losses with an Imposed Normal Curve



The normality assumptions are not valid for the insurance dataset, and nonparametric tests were the appropriate approach to determine statistical associations between variables (Weihs et al., 2018). Figure 2 displays the relative frequencies of incurred losses, claims events, with 93.8% of the 448,755 insurance records having no losses associated with the insurance record with a mean of 331.23 and a standard deviation of 2772.628. A Pearson Chi-Square Test

was used with the large dataset to test the relationships between the binned variables. The Chi-Square Test displayed in Table 8 had an X^2 (117, N = 448,755) = 1037.880, $p < 0.001$, which indicated the relationship between the criminal background variable and the insurance-based credit score variable was statistically significant, and the null hypothesis cannot be rejected. The criminal background variable had 10 categories, and the insurance-based credit score variable had fourteen categories, which made a review of the differences between groups challenging. Because both variables had more than two categories defined, the Chi-Square value is also more complex to interpret, and the effect size should be considered.

Table 8

Pearson Chi-Square Systematic Association Test

Chi-Square Analysis of Criminal Background Events with with Insurance-Based Credit Scores

	Value	*df*	*p* [b]
Pearson Chi-Square	1037.880[a]	117	0.000
Likihood Ratio	1055.252	117	0.000
Linear-by-Linear Association	62.143	1	0.000
N of Valid Cases	448,755		

a. 0 cells (0.00%) have expected count less than 5.
b. Asymptotic Significance (2-sided)
The minimum expected count is 17.44.

While the Chi-Square Test supported a statistically significant association between the two independent variables, the Cramer's V test determined the relationship's effect size. The Cramer's V value displayed in Table 9, where Cramer's V = 0.016, $p < 0.001$, which was closer to zero, indicating the overall effect size was small and suggesting a weaker variable association between the two independent variables (Leech et al., 2015).

Table 9

Cramer's V Strength of Association Test

Nominal by Nominal Symmetric Measures		
N	Cramer's V	p
448,755	0.016	0.000

Hypotheses 2.a.

The third question addressed the model's performance if the criminal background was the leading independent variable with insurance-based credit score removed.

RQ2.a. What is the outcome of the predictive model if the insurance-based score is removed and is replaced with the criminal background variable?

$H2a_o$ = There is no increase in the incremental predictive ability of pure the premium model with the addition of the criminal background data and the removal of the insurance-based score variable.

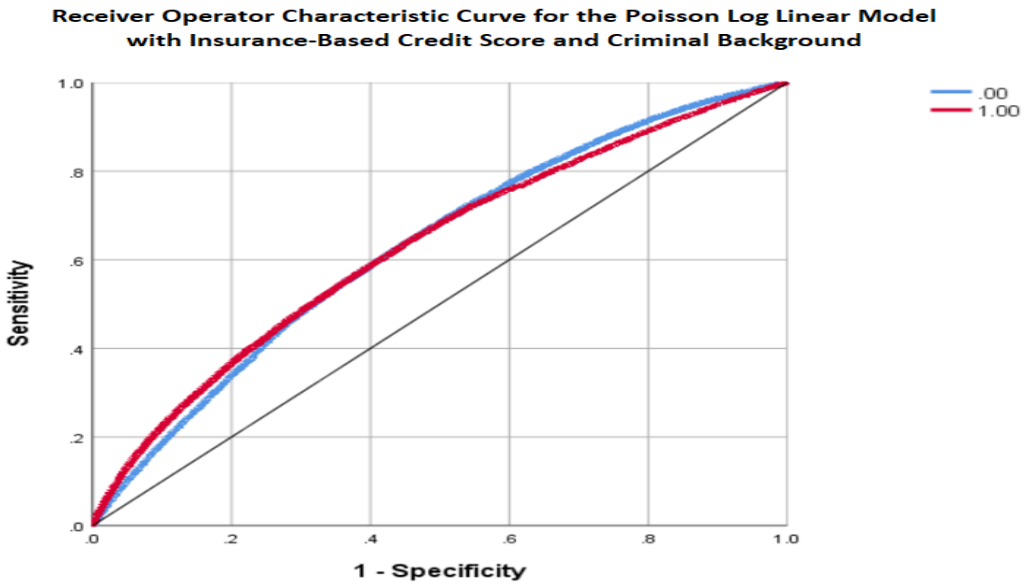
$H2a_1$ = There is an increase in the incremental predictive ability of the pure premium model with the addition of the criminal background data and the removal of the insurance-based score variable .

Table 2 displays the AIC values for the Poisson Regression model's various iterations with the key independent variables. Model 3, with criminal background and without insurance-based credit score, had a lower AIC score 204710.470, $p < 0.001$, than Model 2 with only insurance-based credit score 205159.662, $p < 0.001$, and the null hypothesis can be rejected. The criminal background variable replacement model had a better goodness of fit than with the original independent variables, including insurance-based credit score, and was less complex, requiring a lower number of predictor variables.

Figure 3 displays the Receiver Operator Characteristic (ROC) curve for Model 4, which includes the insurance-based credit score and criminal background variables. The area under the curve (AUC = 0.629) determined how well the model could distinguish between classes. When the two distributions overlap, as shown in Figure 3, the curves indicated there was a 62.9% probability the model will be able to distinguish between positive and negative outcomes for incurred losses. An AUC = 0.500 would indicate the model cannot distinguish between positive and negative outcomes, with the preferred target range for the outcomes to be between 0.700 – 0.800.

Figure 3

Receiver Operator Characteristic Curve for the Poisson Log Linear Model with Insurance-Based Credit Score and Criminal Background



The expectation would be to run the criminal background variable with the full Generalized Linear Model to determine the overall goodness of fit and the probability of accurately predicting the overall pure premium to cover the expected incurred losses.

Relationship of Hypotheses to Research Findings

The research questions and supporting hypotheses were developed to determine if further rate segmentation in the auto insurance industry could potentially decrease rate subsidization through individual driver variables. Research question one and the second part of question two tested the overall goodness of fit for the criminal background variable in relation to the other predictive independent variables and the dependent variable in the predictive model. For the models where the criminal background variable was introduced, there was an improvement in the model's ability to correctly determine pure premium for the independent variables associated with the risk. The Wald Chi-Square outcomes provided additional insights into the relative importance of the variables to the model. Criminal background performed well with a higher Wald Chi-Square result confirming the value of including the variable in the predictive model.

The second research question takes into consideration the relationship between insurance-based credit score and criminal background. While the Chi-Square test was statistically significant for variable association, the effect size was small and weak. The performance of the model variables behaved differently when evaluated independently, and both variables added value to the model. The Receiver Operating Characteristic Curve and the associated Area Under the Curve indicated the model with both variables had a slightly better ability to predict incurred losses, but there was likely room for additional improvement when making future variable adjustments. While the number of predictor variables was the highest in Model 4 displayed in Table 2 when the full Generalized Linear Model is evaluated for product pricing consideration, other variables with less predictive strength can be removed to reduce the model complexity.

Summary of the Findings

The testing results for the research questions concerning the predictive strength of the criminal background variable showed statistical significance for goodness of fit within the Poisson Regression General Linear Models. The lower Akaike Information Criteria (AIC) scores for the models, which included the new variable, were compelling enough and statistically significant, suggesting improved predictive ability when using criminal background to determine individual pure premiums. The Wald Chi-Squared tests also provided additional insights into the criminal background variable's strength when compared with the other top predictive variables within the model. Lass et al. (2016) stated that a condition for determining rate needs during the pricing process was the continuous verification of the current rating structure's adequacy.

For the correlation question, due to the larger number of categories, the crosstabulation data were challenging to interpret effectively even though the matrix meets the expected and observed count criteria (Gloor et al., 2016). While the Chi-Squared test underscored a statistically significant association between criminal background and insurance-based credit scores, the Cramer's *V* test showed the relationship to be small and weak. In the full Generalized Linear model, the variables will be evaluated for predictive strength and overall contribution to the model fit.

Applications to Professional Practice

The research study was intended determine if there were credible ways to decrease rate subsidization in the auto insurance industry. Vehicle insurance is a several billion dollar industry where risk classification practices have significant economic impacts. Abraham (1985) stated risk classification practices also have moral implications because risk sharing is under-analyzed. If insureds have different core risk profiles leading to the development of distinctive loss

experiences, those insureds should be paying different premiums to cover their exposure. The study's findings also addressed the literature gaps, which did not investigate risky individual behavior, partially due to the data not being available for validation.

The most direct application for the study would be to work with the insurance carrier's Data Science Team to order retrospective data on a larger insured population. Consumers sign agreements at the time of application, allowing carriers to order third-party data for rating and underwriting purposes. A larger sample with countrywide data and rating variables brought up to current rate level with the criminal background data appended would provide an actuarially sound approach to pure premium modeling. Kafková and Křivánková (2014) stated the use of Poisson Generalized Linear Models for describing and modeling the data are preferred over more complex models, which could integrate the criminal background into the iterative modeling process. The variable would be incorporated into the model rate order of calculation along with the associated coefficients and filed with the insurance departments. Actuarial exhibits of the variable performance and credibility would accompany the filing.

There will be insurance departments, which will not allow the use of criminal background to determine rates. In those cases, the data can be used to underwrite the risk before allowing the consumer to purchase the insurance. Underwriting criteria allows the carrier to segment risks and protect the book of business by decreasing rate subsidization without increasing a consumer's premium. In certain situations, some consumers who have been convicted of insurance or financial fraud may not be offered a policy from a carrier and will need to be insured under a risk pool plan where premiums are shared across consumers with similar backgrounds at a higher rate. Carriers who ask specific underwriting questions concerning criminal convictions but fail to

validate the answers with third-party data may find their products subject to adverse selection in the future.

From the collection of teachings in the Book of Proverbs, “People learn from one another, just as iron sharpens iron,” and there is a mutual advantage when people and businesses can address uncomfortable situations with an agreed upon solution (Proverbs 27:17). By holding each other accountable for personal behavior and actions, we sharpen one another and become more effective. Using criminal background information to determine accurate premiums will allow carriers to fulfill their agreement to provide fair and accurate rates. They will also expand on the available literature for auto insurance pricing approaches.

Cummings and Tennyson (1992) stated auto insurance was designed to define various financial risk classes and to develop better methods for controlling the risk. The analytics used to drive financial modeling are now the cornerstones for financial and insurance economics (Mankaï & Belgacem, 2016). The more carriers understand the underlying risk through contributing variables like criminal background, the higher the likelihood the risk will be assigned the appropriate premium to cover the exposure. Reducing subsidization helps the market effectively manage insurance premiums and allows for informal financial controls in favor of regulatory mandates.

Recommendations for Action

The supporting statistical analysis to answer the research questions indicated adding the criminal background variable could potentially improve the fit of the current pricing model for the carrier and increase rate segmentation. The full model was not included in the study due to the large number of variables, which would have expanded the analysis’ complexity without adding insights to the criminal background variable’s predictive strength. The results of the

Poisson Log Linear Model iterations and the Chi-Squared outcomes will be shared with the Data Science team. The Data Science team will also want to review the variable relationship between criminal background and insurance-based credit scores to understand the association, while statistically significant, is small and weak. Retrospective data on the existing policies could be ordered from the vendor to append the criminal background data to the larger dataset with more historical policies. The Data Science team will need to run the necessary model validation tests for goodness of fit, measurement of economic lift, and model stability before working with the Product Development team for individual product implementation.

The Product Research and Development team will want to do additional research with the insurance departments for implementation approaches before submitting a rates and rules filing. While there are no current regulations concerning using criminal background for rating, some departments may want additional review time before providing program approval. For individual states, an underwriting rule approach instead of a rating approach may be more appropriate. Programs can run the criminal background reports and offer limited coverage, or no coverage, based on the information returned in an individual's report.

The study will also have implications for sales, marketing, and agent teams who have interactions with consumers. Advanced product information should be provided to allow people to understand the details of the enhanced underwriting or rating process. Individual driver information in its raw format will only be available for the consumer who owns the report. Tailored messaging will support the agents involved in the application process when a report is returned with actionable activity. Criminal background data are considered Fair Credit Reporting Act information safeguards and are in place for proper authorization and consumer protection when using third-party data.

Recommendations for Further Study

The criminal background study intended to determine if additional variables could be identified to further segment risk to reduce rate subsidization in the auto insurance industry. The research addressed the literature gap concerning the use of verified criminal background information to improve the predictive model fit. The analysis also showed an association between the insurance-based credit scores and determined the relationship was not strong enough to remove the criminal background variable from the model.

Years of research and investment in Onboard Diagnostic (OBD) and mobile applications by the insurance industry are being expanded to validate driving behavior and actual miles driven (Baecke & Bocca, 2017; Husnjak et al., 2015; Weidner et al., 2017). The capture of the additional data underscores the desire for consumers to align their insurance premiums with their exposure. Further studies, including data from usage-based insurance records and the driver's criminal background history, may reveal additional characteristics of risky driving behavior, which may also present in the usage based insurance data.

Another actionable area of research would be a correlational analysis between the territory and criminal background variables. Regulatory and consumer advocate groups may require supporting actuarial justification to determine potential bias in the criminal background variable based on urban, suburban, or rural communities. Insurance carriers do not collect data on ethnicity or socioeconomic status, and the analysis could be based on United States Postal Service and census attributes.

The criminal background report can reach back as far seven to ten years, depending on how long the state department of corrections retains events on each individual. Studies on driving violations allow for multiple historical events to be used in both rating and underwriting.

A deeper analysis of the criminal background variable may help determine if tiered rating relativities may be used for multiple violations or if older criminal events should be surcharged differently than more recent criminal activity within the larger rating model.

Reflections

The research study developed during discussions at an advisory board meeting concerning homeowners' insurance fraud. The vendor was open to testing the auto insurance dataset to determine if the criminal background data from the state and federal databases provided insights on loss costs. The one-way analysis supported surcharges for drivers who had criminal convictions in their background. While the translation of someone who exhibits risky behavior in everyday life to driving behavior was not a surprise, the variable's performance in the predictive model was better than anticipated.

The dissertation iterative writing process proved to be more challenging than the actual data analysis. The iterative writing and research process did become part of a daily routine, and at some level, will be missed. The program coursework started within sixty days of the completion of a Master of Science in Predictive Analytics, and some time without a deadline or a weekly progress report will be welcome.

People do question the use of a person's criminal background in the development of insurance rates. Committing a criminal act takes into consideration a series of choices. In Paul's letters to the Romans, people are warned they will be subject to governing authorities, which are not a terror to good conduct, but to bad choices. Insurance carriers are also studying and verifying driving behavior through the use of telematics. The carriers are finding individuals will alter driving behavior when they think they are being watched but will revert to normal behavior within days (Ayuso et al., 2019). People cannot turn risky behavior on and off, and the research

study was able to verify criminal background convictions translate to risky driving behavior statistically.

Summary and Study Conclusions

The quantitative study's objective was to determine if the criminal background variable was predictive of future loss costs, which would help define additional segmentation and reduce premium subsidization. An auto insurance carrier provided a cleansed dataset from 2014-2017 containing rating variables with the criminal background variable appended. A series of Poisson Generalized Linear models were run with various iterations of the criminal background and insurance-based credit score variables to determine if the goodness of fit for the predictive model improved. The model, including both criminal background and insurance-based credit score variables, had the lowest Akaike Information Criteria score providing in-sample prediction error, and was statistically significant. The Data Science team will be investing resources for analyzing the criminal background variable in the full dataset for the next predictive model. While the insurance-based credit score and criminal background had a statistically significant association, the size of the effect was small and weak. The predictive model iterations also supported the use of criminal background alone if insurance-based credit score was not available in for rating.

The analysis and the underwriting report's availability provide an avenue to bridge the literature gap by providing the verified data for modeling. Historically, criminal background data was not available from a third-party vendor and was extraneous in the underwriting process. Introducing a new rating variable incorporates various financial and statistical approaches for developing models to predict more appropriate insurance models. Research to better determine contributing rate segmentation variables should continue with additional modeling and analysis to meet consumer, regulatory, and financial stakeholder expectations.

References

Abraham, K. S. (1985). Efficiency and fairness in insurance risk classification. *Virginia Law Review*, *71*(3), 403-451. https://doi.org/10.2307/1072963

Ahlgrim, K. C., & Jones, J. R. (2014). Insurance rating games: Strike, spares, and bags. *Risk Management and Insurance Review*, *17*(2), 297-313. https://doi.org/10.11118/rmir.12026

Ajiferuke, I., & Famoye, F. (2015). Modelling count response variables in informetric studies: Comparison among count, linear, and lognormal regression models. *Journal of Informetrics*, *9*(3), 499-513. https://doi.org/10.1016/j.joi.2015.05.001

Allison, P. D. (1977). Testing for interaction in multiple regression. *American Journal of Sociology*, *83*(1), 144-153. https://doi.org/10.1086/226510

Andreeva, G., & Matuszyk, A. (2019). The law of equal opportunities or unintended consequences? The effect of unisex risk assessment in consumer credit. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(4), 1467-1998. https://doi.org/10.1111/rssa.12494

Antonio, K., & Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance, Mathematics & Economics*, *40*(1), 59-76. https://doi.org/10.1016/j.insmatheco.2006.02.013

Asamoah, K. (2016). On the credibility of insurance claim frequency: Generalized count models and parametric estimators. *Insurance: Mathematics and Economics*, *70*, 339-353. https://doi.org/10.1016/j.insmatheco.2016.07.003

Aseervatham, V., Lex, C., & Spindler, M. (2016). How do unisex rating regulations affect gender differences in insurance premiums? *The Geneva Papers on Risk and Insurance*, *41*(1), 128-160. https://doi.org/10.10547/gpp.2015.22

Ayuso, M., Guillen, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behavior data. *Transportation*, *46*(3), 735-752. https://doi.org/10.1007/s11116-018-9890-7

Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, *98*, 69-79. https://doi.org/10.1016/j.dss.2017.04.009

Bamattre, S., Hu, R., & Verducci, J. S. (2017). Nonparametric testing for heterogeneous correlation. In A.S. (Eds.), *Big and complex data analysis. Contributions to statistics* (pp. 229-246). Springer. https://doi.org/10.1007/978-3-319-41573-4_12

Barnes, C. M., Dang, C. T., Leavitt, K., Guarana, C., & Uhlmann, E. L. (2018). Archival data in micro-organizational research: A toolkit for moving to a broader set of topics. *Journal of Management*, *44*(4), 1453-1478. https://doi.org/10.1177/0149206315604188

Barraclough, P., af Wåhlberg, A., Freeman, J., Watson, B., & Watson, A. (2016). Predicting crashed using traffic offenses. A meta-analysis that examines potential bias between self-reporting and archival data. *PloS One*, *11*(4), 1-32. https://doi.org/10.1371/journal.pone.0153390

Barth, S., & de Jong, M. D. T. (2017). The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics*, *34*(7), 1038-1058. https://doi.org/10.1016/j.tele.2017.04.013

Bärtl, M., & Krummaker, S. (2020). Prediction of claims in export credit finance: A comparison of four machine learning techniques. *Risks*, *8*(1), 22-29. https://doi.org/10.3390/risks8010022

Baudry, M., & Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry*, *35*(5), 1127-1155. https://doi.org/10.1002/asmb.2455

Baumgartner, C., Gruber, L. F., & Czado, C. (2015). Bayesian total loss estimation using shared random effects. *Insurance, Mathematics & Economics*, *62*, 194-201. https://doi.org/10.1016/j.insmatheco.2015.02.008

Bellatti, J., Brunner, A., Lewis, J., Annadata, P., Eltarjaman, W., Dewri, R., & Thurimella, R. (2017). Driving habits data: Location privacy implications and solutions. *IEEE Security & Privacy*, *15*(1), 12-20. https://doi.org/10.1109/MSP.2017.6

Bermúdez, L., Guillén, M., & Karlis, D. (2018). Allowing for time and cross dependence assumptions between claim counts in ratemaking models. *Insurance, Mathematics & Economics*, *83*, 161-169. https://doi.org/10.1016/j.insmtheco.2018.06.003

Bernard, C., He, X., Yan, J.-A., & Zhou, X. Y. (2015). Optimal insurance design under rank-dependent expected utility. *Mathematical Finance*, *25*(1), 154-186. https://doi.org/10.1111/mafi.12027

Bian, Y., Yang, C., Zhao, L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation Research Part A: Policy and Practice*, *107*, 20-34. https://doi.org/10.1016/j.tra.2017.10.018

Binder, S., & Muβhoff, J. (2017). *Global insurance industry insights: An in-depth perspective* (7th ed.). McKinsey & Company.

Blesa, A., Íñiguez, D., Moreno, R., & Ruiz, G. (2020). Use of open data to improve automobile insurance premium rating. *International Journal of Market Research: The Journal of the Market Research Society*, *62*(1), 58-78. https://doi.org/10.1177/1470785319862734

Bloomfield, J., & Fisher, M. (2019). Quantitative research design. *JARNA: Official Journal of the Australasian Rehabilitation Nurses' Association*, *22*(2), 27-30. https://doi.org/10.33235/jarna.22.2.27-30

Bolancé, C., Guillen, M., Nielson, J., & Thuring, F. (2018). Price and profit optimization for financial services. *Risks*, *6*(1), 9-21. https://doi.org/10.3390/risks6010009

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431-449. https://doi.org/10.1037/a0038047

Boyer, M. M., & Owadally, I. (2015). Underwriting apophenia and cyptics: Are cycles statistical figments of our imagination? *Geneva Papers on Risk & Insurance*, *40*(2), 232-255. https://doi.org/10.1057/gpp.2014.12

Brannon, J. (2016). *Mixing methods: Qualitative and quantitative research.* Routledge.

Brockett, P. L., & Golden, L. L. (2007). Biological and psychobehavioral correlates of credit score and automobile insurance losses: Toward an explication of why credit scoring works. *Journal of Risk and Insurance*, *74*(1), 23-63. https://doi.org/10.1111/j.1539-6975.2007.002021.x

Brown, R. L., Charters, D., Gunz, S., & Haddow, N. (2007). Colliding interests – Age as an automobile insurance rating variable: Equitable rate-making or unfair discrimination? *Journal of Business Ethics*, *72*(2), 103-114. https://doi.org/10.1007/s10551-006-9160-z

Carmichael, L., Stalla-Bourdillion, S., & Staab, S. (2016). Data mining and automated discrimination: A mixed legal/technical perspective. *IEEE Intelligent Systems*, *31*(6), 51-55. https://doi.org/10.1109/MIS.2016.96

Castignani, G., Derrmann, T., Frank, R., & Engel, T. (2015). Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. *IEEE Intelligent Transportation Systems Magazine*, *7*(1), 91-102. https://doi.org/10.1109/MITS.2014.2328673

Casualty Actuarial Society. (1990). *Foundations of Casualty Actuarial Science.* Casualty Actuarial Society.

Cather, D. A. (2018). Cream skimming: Innovations in insurance risk classification and adverse selection. *Risk Management and Insurance Review*, *21*(2), 335-366. https://doi.org/10.1111/rmir.12012

Charpentier, A. (2017). Computational actuarial science with r. *Journal of Risk and Insurance*, *84*(1), 267-270. https://doi.org/10.1111/jori.12185

Choi, S., Hardigree, D., & Thistle, P. D. (2002). The property/liability insurance cycle: A comparison of alternative models. *Southern Economic Journal*, *68*(3), 530-548. https://doi.org/10.2307/1061716

Chowdhury, S., Chatterjee, S., Mallick, H., Banerjee, P., & Garai, B. (2019). Group regularization for zero-inflated Poisson regression model with application to insurance ratemaking. *Journal of Applied Statistics*, *46*(9), 1360-0532. https://doi.org/10.1080/0266763.2018.1555232

Cohen, A., & Siegelman, P. (2010). Testing for adverse selection in insurance markets. *The Journal of Risk and Insurance*, *77*(1), 39-84. https://doi.org/10.1111/j.1539-6975.2009.01337.x

Cohen, J. (2015). Studying law studying surveillance. *Surveillance & Society*, *13*(1), 91-101. https://doi.org/10.24908/ss.v13i1.5160

Cole, C., Maroney, P., McCullough, K., & Powell, L. (2015). Automobile insurance vehicle repair practices: Politics, economics, and consumer interests. *Risk Management and Insurance Review*, *18*(1), 101-128. https://doi.org/10.1111/rmir.12032

Collins, A., Joseph, D., & Bielaczyc, K. (2009). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, *13*(1), 15-42. https://doi.org/10.1207/s15327809jls.1301-2

Cotteleer, M. J., & Wan, X. (2016). Does the starting point matter? The literature-driven and the phenomenon-driven approaches of using corporate archival data in academic research. *Journal of Business Logistics*, *37*(1), 26-33. https://doi.org/10.1111/jbl.12144

Cova, B., Gaglio, G., Weber, J., & Chanial, P. (2016). Organizational sensemaking of non-ethical consumer behavior: Case study of a French mutual insurance company. *Journal of Business Ethics*, *148*(4), 783-799. https://doi.org/10.1077/s10551-016-3102-1

Creswell, J., & Creswell, J. (2018). *Research design: Qualitative, quantitative, and mixed method approaches* (5th ed.). Sage.

Cummins, J., & Tennyson, S. (1992). Controlling automobile insurance costs. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *6*(2), 95-115. https://doi.org/10.1257/jep.6.2.95

Cummins, J. D., & Xie, X. (2016). Efficiency and productivity in the US property-casualty liability insurance industry: Ownership, structure, product and distribution strategies. In J. Zhu (Eds.), *Data envelopment analysis. International series in operations research & management science*, *Vol. 238* (pp. 113-163). https://doi.org/doi: 10.1007/978-1-4899-7684-0_6

Curtis, E. A., Comiskey, C., & Dempsey, O. (2016). Importance and use of correlational research. *Nurse Researcher*, *23*(6), 20-25. https://doi.org/10.7747/nr.2016.e1382

David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, *20*, 147-156. https://doi.org/10.1016/S2212-5671(15)00059-3

Davoudi Kakhki, F., Freeman, S., & Mosher, G. (2018). Analyzing large workers' compensation claims using generalized linear models and Monte Carlo simulation. *Safety*, *4*(4), 57-69. https://doi.org/10.3390/safety4040057

Denuit, M., & Trufin, J. (2018). Collective loss reserving with two types of claims in motor third party liability insurance. *Journal of Computational and Applied Mathematics*, *335*, 168-184. https://doi.org/10.1016/j.cam.2017.11.044

Derring, R. A., Johnston, D. J., & Sprinkel, E. A. (2006). Auto insurance fraud: Measurements and efforts to combat it. *Risk Management and Insurance Review*, *9*(2), 109-130. https://doi.org/10.111/j.1540-6296.2006.00089x

Dijksterhuis, C., Lewis-Evans, B., Jelijs, B., de Waard, D., Brookehuis, K., & Tucha, O. (2015). The impact of immediate or delayed feedback on driving behavior in a simulated pay-as-you-drive system. *Accident Analysis and Prevention*, *75*, 93-104. https://doi.org/10.1016/j.aap.2014.11.017

Dionne, G., & Vanasse, C. (1992). Automobile insurance ratemaking in the presence of asymmetrical information. *Journal of Applied Econometrics*, *7*(2), 149-165. https://doi.org/10.1002/jae.3950070204

Duan, Z., Chang, Y., Wang, Q., Chen, T., & Zhao, Q. (2018). A logistic regression based auto insurance rate-making model designed for insurance rate reform. *International Journal of Financial Studies*, *6*(1), 1-17. https://doi.org/10.3390/ijfs6010018

Edmonds, W. A., & Kennedy, T. D. (2017). *An applied guide to research designs: Quantitative, qualitative and mixed method* (2nd ed.). Sage.

Einav, L., Finkelstein, A., Kluender, R., & Schrimpt, P. (2016). Beyond statistics: The economic content of risk scores. *American Economic Journal: Applied Economics*, *8*(2), 195-224. https://doi.org/10.1257/app.20150131

Eling, M., & Pankoke, D. (2016). Costs and benefits of financial regulation: An empirical assessment for insurance companies. *The Geneva Papers on Risk and Insurance*, *41*(4), 529-554. https://doi.org/10.1057/gpp.2016.11

Eryilmaz, S. (2016). Compound Markov negative binomial distribution. *Journal of Computational and Applied Mathematics*, *292*, 1-6. https://doi.org/10.1057/10.1016/j.com.2015.06.026

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics*, *5*(1), 1-4. https://doi.org/10.11648/j.ajtas.20160501.11

Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using big data analytics: A case study of the insurance industry. *Computers & Industrial Engineering*, *101*, 554-564. https://doi.org/10.1016/j.cie.2016.09.011

Favaretto, M., De Clercq, E., & Elger, B. S. (2019). Big data and discrimination: perils, promised and solutions. A systematic review. *Journal of Big Data*, *6*(1), 1-27. https://doi.org/10.1186/s40537-019-0177-4

Feest, J. (1968). Compliance with legal regulation: Observation of stop sign behavior. *Law & Society Review*, *2*, 447-462. https://doi.org/10.2307/3052898

Fox, J. (2016). *Applied regression analysis and generalized linear models.* Sage.

Frees, E. W., Derrig, R. A., & Meyers, G. (2014). *Predictive modeling applications in actuarial science* (Vol. 1). Cambridge University Press.

Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate frequency-severity regression models in insurance. *Risks*, *4*(1), 1-36. https://doi.org/10.3390/risks4010004

Frees, E. W., Meyers, G., & Cummings, A. D. (2014). Insurance ratemaking and a Gini index. *The Journal of Risk and Insurance*, *81*(2), 335-366. https://doi.org/10.1111/j.1539-6975.2012.01507x

Frezal, S., & Barry. L. (2019). Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 1-10. https://doi.org/10.1007/s10551-019-04171-2

Fuzi, M. F. M., Jemain, A. A., & Ismail, N. (2016). Bayesian quantile regression model for claim count data. *Insurance, Mathematics & Economics*, *66*, 124-137. https://doi.org/10.1016/j.insmatheco.2015.11.004

Gaffney, C., & Ben-Israel, A. (2016). A simple insurance model: Optimal coverage and deductible. *Annals of Operations Research*, *237*(1-2), 263-279. https://doi.org/10.1007/s10479-013-1469-2

Gan, G., & Valdez, E. A. (2018). Fat-tailed regression modeling with spliced distributions. *North American Actuarial Journal*, *22*(4), 554-573. https://doi.org/10.1080/10920277.2018.1462718

Gao, J., Xie, C., & Tao, C. (2016, March). Big Data Validation and Quality Assurance--Issues, Challenges, and Needs. In *2016 IEEE symposium on service-oriented system engineering* (SOSE) (pp. 433-441). IEEE. https://doi.org/10.1109/SOSE.2016.63

Garfunkel, S. L. (2015). De-identification of personal information. *National Institute of Standards and Technology Internal Report 8053.* https://doi.org/10.6028/NIST.IR.8053

Garrido, J., Genest, C., & Schulz, J. (2016). Generalized linear models for dependent frequency and severity of insurance claims. *Insurance, Mathematics & Economics*, *70*, 205-215. https://doi.org/10.1016/j.insmathco.2016.06.006

Gennaioli, N., Shleifer, A., & Vishny, R. (2015). Money doctors. *The Journal of Finance*, *70*(1), 91-114. https://doi.org/10.1111/jofi.12188

Gloor, G. B., Macklaim, J. M., & Fernandes, A. D. (2016). Displaying variation in large datasets: Plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *25*(3), 971-979. https://doi.org/10.1080/10618600.2015.1131161

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2020). *Generalized linear models for insurance rating* (2nd ed.). The Casualty Actuarial Society.

Golden, L. L., Brockett, P. L., Ai, J., & Kellison, B. (2016). Empirical evidence on the use of credit scoring for predicting insurance losses with psycho-social and biochemical explanations. *North American Actuarial Journal*, *20*(3), 233-251. https://doi.org/10.1080/10920277.2016.1209118

Gómez-Déniz, E. (2008). A generalization of credibility theory obtained by using the weighted average loss function. *Insurance, Mathematics & Economics*, *42*(2), 850-854. https://doi.org/10.1016/j.insmatheco.2007.09.002

Gómez-Déniz, E. (2016). Bivariate credibility bonus-malus premiums distinguishing between two types of claims. *Insurance, Mathematics & Economics*, *70*(1), 117-124. https://doi.org/10.1016/j.insmatheco.2016.06.009

Graham, L., & Xie, X. (2007). The United States insurance market: Characteristics and trends. In D. Cummings & B. Vernard (Eds.), *Handbook of international insurance* (pp. 25-120). Springer.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659-3667. https://doi.org/10.1016/j.eswa.2011.09.058

Guelman, L., & Guillén, M. (2014). A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications*, *41*(2), 387-396. https://doi.org/10.1016/j.eswa.2013.07.059

Haberman, S., & Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of Royal Statistical Society*, *45*(4), 407-436. https://doi.org/10.2307/2988543

Harrington, S. (1986). Estimation and testing for functional form in pure premium regression models. *ASTIN Bulletin*, *16*(1), S31-S43. https://doi.org/10.1017/S0515036100011636

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*(3), 66-67. https://doi.org/10.1136/eb-2015-102129

Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. *Scandinavian Actuarial Journal*, *2018*(8), 681-705. https://doi.org/10.1080/03461238.2018.1429300

Hinterhuber, A., & Liouzu, S. M. (2017). Is innovation in pricing your next source of competitive advantage? In A. Hinterhuber & S.M. Liouzu (Eds.), *Innovation in pricing: Contemporary theory and best practices* (pp. 11-27). Routledge. https://doi.org/10.4324/9781315184845-2

Honka, E. (2014). Quantifying search and switching costs in the US auto insurance industry. *The Rand Journal of Economics*, *45*(4), 847-884. https://doi.org/10.1111/1756-2171.12073

Hoy, M. (1982). Categorizing risks in the insurance industry. *The Quarterly Journal of Economics*, *97*(2), 321-336. https://doi.org/10.2307/1880761

Hsu, Y. C., Shui, Y. M., Chou, P. L., & Chen, Y. M. J. (2015). Vehicle insurance and the risk of road traffic accidents. *Transportation Research*, *74*, 201-206. https://doi.org/10.1016/j.ta.2015.02.015

Hua, L. (2015). Tail negative dependence and its applications for aggregate loss modeling. *Insurance: Mathematics and Economics*, *61*, 135-145. https://doi.org/10.1016/j.insmatheco.2015.01.001

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, *100*, 816-825. https://doi.org/10.1016/j.proeng.2015.1.436

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social and biomedical sciences: An introduction.* Cambridge University Press.

Insurance Information Institute. (2019). *Background on: Credit scoring.* Insurance Information Institute, Inc.

Ippolito, R. A. (1979). The effects of price regulation in the automobile insurance industry. *The Journal of Law & Economics*, *22*(1), 55-89. https://doi.org/10.1086/466933

Ishida, C., Chang, W., & Taylor, S. (2016). Moral intensity, moral awareness and ethical predispositions: The case of insurance fraud. *Journal of Financial Services Marketing*, *21*(1), 4-18. https://doi.org/10.1057/fsm.2015.26

Isotupa, K. P. S., Kelly, M., & Kleffner, A. (2019). Experience-rating for high-risk, low-risk, and novice drivers. *North American Actuarial Journal*, *23*(3), 395-411. https://doi.org/10.1080/10920277.2019.1572524

Jaspersen, J. G., & Richter, A. (2015). The wealth effects of premium subsidies on moral hazard in insurance markets. *European Economic Review*, *77*, 139-153. https://doi.org/10.1016/j.euroecorev.2015.03.007

Jee, B. (1989). A comparative analysis of alternative pure premium models in the automobile risk classification system. *The Journal of Risk and Insurance*, *56*(3), 434-459. https://doi.org/10.2307/253167

Jeong, H., Valdez, E. A., Ahn, J. Y., & Park, S. (2017). Generalized linear mixed models for dependent compound risk models. *SSRN*, 1-23. https://doi.org/10.2139/ssrn.3045360

Jeong, H., Gan, G., & Valdez, E. (2018). Association rules for understanding policyholder lapses. *Risks*, *6*(3), 69-87. https://doi.org/10.3390/risks6030069

Johnson, K. A., Hook, J. N., Davis, D. E., Van Tongeren, D. R., Sandage, S. J., & Crabtree, S. A. (2016). Moral foundation priorities reflect U.S. Christian's individual difference in religiosity. *Personality and Individual Differences*, *100*, 55-61. https://doi.org/10.1016/j.paid.2015.12.037

Joksch, H. C. (1981). A relation between collision claim frequency and driver age distribution in HLDI data. *Accident Analysis & Prevention*, *13*(2), 147-151. https://doi.org/10.1016/0001-4575(81)90026-9

Kadiyala, S., & Heaton, P. S. (2017). The effect of health insurance coverage expansions on auto liability claims and costs. *Rand Working Paper Series*, 1-35. https://doi.org/10.2139/ssrn.3087503

Kafková, S., & Křivánková, L. (2014). Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silvculturae Mendelianae Brunensis*, *62*(2), 383-388. https://doi.org/10.11118/actuaun20146202383

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technological Law Journal*, *34*, 190-217. https://doi.org/10.15779/Z38TD9N83J

Kang, S., & Song, J. (2018). Feature selection for continuous aggregate response and its application to auto insurance data. *Expert Systems with Applications*, *93*(1), 104-117. https://doi.org/10.1016/j.eswa.2017.10.007

Karmila, S., Nurrohmah, S., & Sari, S. F. (2020). Claim reserve prediction using credibility theory for chain ladder method. *Journal of Physics: Conference Series*, *1442*(1), 1-6. https://doi.org/10.1088/1742-6596/1442/1/012038

Kelly, M., & Nielson, N. (2006). Age a variable in insurance pricing and risk classification. *The Geneva Papers on Risk and Insurance*, *31*(2), 212-232. https://doi.org/10.1057/palgrave.gpp.2510082

Khashei, M., Zeinal, H., & Ali Bijari, J. (2012). A novel hybrid classification model of artificial neural networks and multiple linear regression models. *Expert Systems with Applications*, *39*(3), 2606-2620. https://doi.org/10.1016/j.eswa.2011.08.116

Kim, S. K., & Min, S. (2015). Business model innovation performance: When does adding a new business model benefit an incumbent? *Strategic Entrepreneurship*, *9*(1), 34-57. https://doi.org/10.1002/sej.1193

Kimball, S. L., & Boyce, R. N. (1958). The adequacy of state insurance rate regulation: The Mccarran-Ferguson Act in historical perspective. *Michigan Law Review*, *56*(4), 545-578. https://doi.org/10.2307/1286053

Kimura, K., & Waki, H. (2018). Minimization of Akaike's information criterion in linear regression analysis via mix integer nonlinear program. *Optimization Methods & Software*, *33*(3), 633-649. https://doi.org/10.1080/10556788.2017.1333611

Kiviat, B. (2019). The moral limits of predictive practices: The case of credit-based insurance scores. *American Sociological Review*, *84*(6), 1134-1158. https://doi.org/10.1177/000312219884917

Klieštik, T., Kočišovà, K., & Mišankovà, M. (2015). Logit and Probit model used for prediction of financial health of company. *Procedia Economics and Finance*, *23*, 850-855. https://doi.org/10.1016/S2212-5671(15)00485-2

Kofman, P., & Nini, G. P. (2013). Do insurance companies possess an information monopoly? Empirical evidence from auto insurance. *The Journal of Risk and Insurance*, *80*(4), 1001-1026. https://doi.org/10.1111/j.1539-6975-2012.01487.x

Krippner, G. R. (2017). Democracy of credit: Ownership and the politics of credit access in late twentieth-century America. *American Journal of Sociology*, *123*(1), 1-47. https://doi.org/10.1086/692274

Krzywinski, M., & Altman, N. (2015). Multiple linear regression. *Nature Methods*, *12*(12), 1103-1104. https://doi.org/10.38/nmeth.3665

Kurlychek, M. C., Brame, R., & Bushway, S. D. (2006). Scarlett letters and recidivism: Does an old criminal record predict future offending? *Criminology and Public Policy*, *5*(3), 483-503. https://doi.org/10.1111/j.1745-9133-2066.00397.x

Lacy, W. (2017). Use and impact of credit in personal lines insurance premiums pursuant to Ark. Code Ann. §23-67-415. *A Report to the Legislative Council and the Senate and House Committees on Insurance and Commerce or the Arkansas General Assembly (As required by Act 1452 of 2003).*Arkansas Insurance Department.

Landes, X. (2015). How fair is actuarial fairness? *Journal of Business Ethics*, *128*(3), 519-533. https://doi.org/10.1007/s10551-014-21210-0

Lass, D., Scheister, H., & Wagner, J. (2016). Empirical findings on motor insurance pricing in Germany, Austria and Switzerland. *The Geneva Papers on Risk and Insurance*, *41*(3), 398-431. https://doi.org/10.1057/gpp.2015.30

Lee, D., Sorich, S., Miller, M. J., & Schneider, R. (2005). Give us some credit: The use of credit information in insurance underwriting and rating. *Risk Management and Insurance Review*, *8*(1), 31-41. https://doi.org/1540-6296.2005.00048.x

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2015). *IBM SPSS for intermediate statistics: Use and interpretation* (5th ed.). Routledge.

Lehtonen, T. K., & Liukko, J. (2015). Producing solidarity, inequality and exclusion through insurance. *Res Publica*, *21*(2), 155-169. https://doi.org/10.1007/s11158-015-9270-5

Lemaire, J., Park, S. C., & Wang, K. C. (2015). The impact of covariates on a bonus-malus system: An application of Taylor's model. *European Actuarial Journal*, *5*(1), 1-10. https://doi.org/10.1007/s13385-015-0107-6

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the Association for Computing Machinery*, *61*(10), 36-43. https://doi.org/10.1145/3233231

Little, T. D. (2013). *The Oxford handbook of quantitative methods*: *Statistical analysis* (Vol. 2). Oxford University Press.

Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *The Journal of the Operational Research Society*, *56*(9), 1099-1108. https://doi.org/10.1057/palgrave.jors.2601976

Livieris, I., Kiriakidou, N., Kanavos, A., Tampakas, V., & Pintelas, P. (2018). On ensemble SSL algorithms for credit scoring problem. *Informatics*, *5*(4), 1-16. https://doi.org/10.3390/informatics5040040

Lock, I., & Seele, P. (2015). Quantitative content analysis as a method for business ethics research. *Business Ethics*, *24*, S24-S40. https://doi.org/10.1111/beer.12095

Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2016). *Statistics: Unlocking the power of data.* Wiley Custom Learning Solutions.

Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y.-C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Part A: Policy and Practice*, *113*, 243-258. https://doi.org/10.1016/j.tra.2018.04.013

Mankaï, S., & Belgacem, A. (2016). Interactions between risk taking, capital, and reinsurance for property liability insurance firms. *The Journal of Risk and Insurance*, *83*(4), 1007-1043. https://doi.org/10.1111/jori.12080

McCusker, K., & Gunaydin, S. (2015). Research using qualitative, quantitative or mixed methods and choice based on the research. *Perfusion*, *30*(7), 537-542. https://doi.org/10.1177/0267659114559116

Meyers, G., & Van Hoyweghen, I. (2018). Enacting actuarial fairness in insurance: From fair discrimination to behavior-based fairness. *Science as Culture*, *27*(4), 413-438. https://doi.org/10.1080/09505431.2017.1398223

Mikoski, P., Zlupko, G., & Owens, A. (2019). Drivers' assessments of the risks of distraction, poor visibility at night, and safety-related behaviors of themselves and other drivers. *Transportation Research Part F: Traffic Psychology and Behavior*, *62*, 416-434. https://doi.org/10.1016/j.trf.2019.01.011

Miljkovic, T., & Fernândez, D. (2018). On two mixture-based clustering approached used in modeling an insurance portfolio. *Risks*, *6*(2), 1-18. https://doi.org/10.3390/risks6020057

Miljkovic, T., & Grün, B. (2016). Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, *70*, 387-396. https://doi.org/10.1016/j.insmatheco.2016.06.019

Morata, L. B. I. (2009). A priori ratemaking using bivariate Poisson regression models. *Insurance: Mathematics and Economics*, *44*(1), 135-141. https://doi.org/10.1016/j.insmatheco.2008.11.005

Morgan, D. L. (2018). Living within blurry boundaries: The value of distinguishing between qualitative and quantitative research. *Journal of Mixed Methods Research*, *12*(3), 268-279. https://doi.org/10.1177/1558689816686433

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2013). *IBM SPSS for introductory statistics: Use and interpretation* (5th ed.). Routledge.

Moro, S., Cortez, P., & Rita, P. (2017). A framework for increasing the value of predictive data-driven model by enriching domain characterization with novel features. *Neural Computing & Applications*, *28*(6), 1515-1523. https://doi.org/10.1007/s00521-015-2157-8

Morris, D. S., Schwarcz, D., & Teitelbaum, J. C. (2017). Do credit-based insurance scores proxy for income in predicting auto claim risk? *Journal of Empirical Legal Studies*, *14*(2), 397-423. https://doi.org/10.1111/jells.12151

Nagrecha, S., Johnson, R. A., & Chawla, N. V. (2018). FraudBuster: Reducing fraud in an auto insurance market. *Big Data*, *6*(1), 3-12. https://doi.org/10.1089/big.2017.0083

Najafabadi, A. T. P., Atatalab, F., & Najafabadi, M. O. (2017). Credibility premium to rate-making systems. *Communications in Statistics*, *46*(1), 415-426. https://doi.org/10.1080/03610926.2014.995823

National Association of Insurance Commissioners. (2019). *2018 Market Share Reports for Property/Casualty Groups and Companies by State and Countrywide.* National Association of Insurance Commissioners.

Nicholson, J. E. (2019). Challenges for the insurance industry in the future. *Journal of Insurance Regulation*, *38*(6), 1-26. https://doi.org/10.2139/ssrn.3218082

O'Brien, J. (2018). Age, autos, and the value of a statistical life. *Journal of Risk and Uncertainty*, *57*(1), 51-79. https://doi.org/10.1007/s11166-018-9285-3

Ohlsson, J., Händel, P., Han, S., Welch, R., vom Brocke, J., & Schmiedel, T. (2015). Process innovation with disruptive technology: Lessons learned from a smartphone-based insurance telematics initiative. In J. vom Brocke & T. Schmiedel (Eds.), *BPM – Driving innovation in a digital world* (pp. 85-101). Springer. https://doi.org/10.1007/978-3-319-14430-6_7

Omari, C. O., Nyambura, S. G., & Mwanqi, J. M. W. (2018). Modeling the frequency and severity of auto insurance claims using statistical distributions. *Journal of Mathematical Finance*, *8*(1), 137-160. https://doi.org/10.4236/jmf.2018.81012

Palumbo, A. J., Pfeiffer, M. R., Metzger, K. B., & Curry, A. E. (2019). Driver licensing, motor-vehicle crashes, and moving violation among older adults. *Journal of Safety Research*, *71*, 87-93. https://doi.org/10.1016/j.jsr.2019.09.019

Park, S. C., Kim, J. H. T., & Ahn, J. Y. (2018). Does hunger for bonuses driver the dependence between claim frequency and severity? *Insurance, Mathematics & Economics*, *83*, 32-46. https://doi.org/10.1016/j.insmatheco.2018.09.002

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Payandeh, N., & Amir, T. (2010). A new approach to the credibility formula. *Insurance, Mathematics & Economics*, *46*(2), 334-338. https://doi.org/10.1016/10.1016/j.insmatheco.2009.11.007

Pechon, F., Denuit, M., & Trufin, J. (2019). Multivariate modeling of multiple guarantees in motor insurance of a household. *European Actuarial Journal*, *9*(2), 595-602. https://doi.org/10.1007/s13385-019-00201-5

Pechon, F., Trufin, J., & Denuit, M. (2018). Multivariate modeling of household claim frequencies in motor third-party liability insurance. *ASTIN Bulletin*, *48*(3), 969-993. https://doi.org/10.1017/asb.2018.21

Porrini, D. (2015). Risk classification efficiency and the insurance market regulation. *Risks*, *3*(4), 445-454. https://doi.org/10.3390/risks3040445

Qian, W., Yang, Y., & Zou, H. (2016). Tweedie's compound Poisson model with grouped elastic net. *Journal of Computational and Graphical Statistics*, *25*(2), 606-625. https://doi.org/10.1080/10618600.2015.1005213

Quan, Z., & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), 377-407. https://doi.org/10.1515/demo-2018-0022

Quijano-Xacur, O. A., & Garrido, J. (2015). Generalized linear models for aggregate claims: To Tweedie or not? *European Actuarial Journal*, *5*(1), 181-202. https://doi.org/10.1007/s13385-015-0108-5

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research*, *8*(3), 148-151. https://doi.org/10.4103/picr.PICR_87_17

Richman, R. (2018). AI in actuarial science. *SSRN*, 1-53. https://doi.org/10.2139/ssrn.3218082

Roy, R., & George, K. T. (2017). Detecting insurance claims fraud using machine learning techniques. *2017 International Conference on Circuit, Power, and Computing Technologies (ICCPCT)*, 1-6. https://doi.org/10.1109/ICCPCT.2017.8074258

Salkind, N. J., & Frey, B. B. (2020). *Statistics for people who (think they) hate statistics.* Sage.

Samii, C. (2016). Causal empiricism in quantitative research. *Journal of Politics*, *78*(3), 941-955. https://doi.org/10.1086/686690

Schwarcz, D. (2018). Ending public utility style rate regulation in insurance. *Yale Journal on Regulation*, *35*, 941-990.

Segovia-Vargas, M. J., Camacho-Miñano, M. D. M., & Pascual-Ezama, D. (2015). Risk factors selection in automobile insurance policies: A way to improve the bottom line of insurance companies. *Revista brasileira de gestão de negócios*, *15*(57), 1228-1245. https://doi.org/10.7819/rbgn.v17i57.1741

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency-severity modeling of insurance claims. *Insurance: Mathematics and Economics*, *64*, 417-428. https://doi.org/10.1016/j.insmatheco.2015.07.006

Shim, J. (2017). An investigation of market concentration and financial stability in property-liability insurance industry. *Journal of Risk and Insurance*, *84*(2), 567-597. https://doi.org/10.1111/jori.12091

Siedlecki, S. L. (2020). Understanding descriptive research designs and methods. *Clinical Nurse Specialist*, *34*(1), 8-12. https://doi.org/10.1097/NUR.0000000000000493

Soleymanian, M., Weinberg, C. B., & Zhu, T. (2019). Sensor data and behavior tracking: does usage-based auto insurance benefit drivers? *Marketing Science*, *38*(1), 21-43. https://doi.org/10.1287/mksc.2018.1126

Spedicato, G. A., Dutang, C., & Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMS. *Variance, Casualty Actuarial Society*, *12*(1), 69-89. https://doi.org/hal-01942038

Spinnewijn, J. (2017). Heterogeneity, demand for insurance and adverse selection. *American Economic Journal: Economic Policy*, *9*(1), 308-343. https://doi.org/10.1257/pol.20140254

Stiff, C., Bell, R. L., & Tandon, S. (2019). Credit-based insurance scores and the cost to consumers. *Journal of Accounting and Finance*, *19*(8), 143-149. https://doi.org/10.33423/jaf.v19i8.2621

Störmer, T. (2015). Optimizing insurance pricing by incorporating consumers' perceptions of risk classification. *Zeitschrift für die Gesamte Versicherungs-wissenschaft*, *104*(1), 11-37. https://doi.org/10.1007/s12297-014-0287-1

Tan, C. I. (2016a). Optimal design of a bonus-malus system: Linear relativities revisited. *Annals of Actuarial Science*, *10*(1), 52-64. https://doi.org/10.1017/S1748499515000111m

Tan, C. I. (2016b). Varying transition rules in bonus-malus systems: From rules specification to determination of optimal relativities. *Insurance, Mathematics & Economics*, *68*, 134-140. https://doi.org/10.1016/j.insmatheco.2016.03.007

Tan, C. I., Li, J., Siu-Hang, J., & Balasooriya, U. (2015). Optimal relativities and transition rules of a bonus-malus system. *Insurance, Mathematics & Economics*, *61*, 255-263. https://doi.org/10.1016/j.insmatheco.2015.02.001

Tayal, D. K., Jain, A., Arora, S., Agarwal, S., Gupta, T., & Tyagi, N. (2015). Crime detection and criminal identification in India using data mining techniques. *AI & Society*, *30*(1), 117-127. https://doi.org/10.1007/s00146-014-0539-6

Thiery, Y., & Schoubroeck, C. (2006). Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance*, *31*(2), 190-211. https://doi.org/10.1057/palgrave.gpp.2510078

Tzougas, G., Hoon, W. L., & Lim, J. M. (2019). The negative binomial-inverse Gaussian regression model with an application to insurance ratemaking. *European Actuarial Journal*, *9*(1), 323-344. https://doi.org/10.1007/s13385-018-0186-2

Uraschi, G., Horodnic, I. A., & Zait, A. (2015). How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Economics and Finance*, *20*, 679-686. https://doi.org/10.1016/S2212-5671(15)00123-9

Valecky, J. (2016). Modelling claim frequency in vehicle insurance. *Acta Universitatis Agriculturae et Silvculturae Mendelianae Brunensis*, *64*(2), 683-689. https://doi.org/10.11118/actaun201664020683

van Rijnsoever, F. J. (2017). (I can't get no) saturation: A simulation and guidelines for sample sizes in qualitative research. *PLoS One*, *12*(7), 1-17. https//doi.org/10.1371/journal.pone.0181689

Vavouranakis, P., Panagiotakis, S., & Mavormoutakis, C. X. (2017). Smartphone-based telematics for usage based insurance. In C.X. Mavormoutakis, G. Mastorakis, & C. Dobre (Eds.), *Advances in mobile cloud computing and big data in the 5G era. Studies in big data* (Vol. 22; pp. 309-339). Springer. https://doi.org/10.1007/978-3-319-45145-9_13

Vinsel, L. (2019). *Moving violations: Automobiles, experts and regulations in the United States.* Johns Hopkins University Press.

Watson, R. (2015). Quantitative research. *Nursing Standard*, *29*(31), 44-48. https://doi.org/ns.29.31.44.e8681

Weidner, W., Transchel, F. W. G., & Weidner, R. (2017). Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science*, *11*(2), 213-236. https://doi.org/10.1017/S748499516000130

Weihs, L., Drton, M., & Meinshausn, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika*, *105*(3), 547-562. https://doi.org/10.1093/biomet/asy021

Weiss, M. A., & Choi, B. P. (2008). State regulation and the structure, conduct, efficiency and performance of us auto insurers. *Journal of Banking and Finance*, *32*(1), 134-156. https://doi.org/10.106/j.bankfin.2007.09.010

Weiss, M. A., Tennyson, S., & Regan, L. (2010). The effects of regulated premium subsidies on insurance costs: An empirical analysis of automobile insurance. *Journal of Risk and Insurance*, *77*(3), 597-624. https://doi.org/10.1111/j.1539-6975.2010.01360.x

Werner, G., & Modlin, C. (2016). *Basic ratemaking* (5th ed.). Casualty Actuarial Society.

Witt, R. C. (1974). Credibility standards and pricing in automobile insurance. *American Risk and Insurance*, *41*(3), 375-396. https://doi.org/10.2307/252041

Wüthrich, M. V. (2019). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal*, 1-24. https://doi.org/10.1007/s13385-019-00215-z

Wüthrich, M. V., & Buser, C. (2019). Data analytics for non-life insurance pricing. *Swiss Finance Institute Research Paper No. 16-68.* https://doi.org/10.2139/ssrn.2870308

Xie, S., & Lawniczak, A. (2018). Estimating major risk factor relativities in rate filing using generalized linear models. *International Journal of Financial Studies*, *6*(4), 84-99. https://doi.org/10.3390/ijfs6040084

Xie, Y.-T., Li, Z.-X., & Parsa, R. A. (2018). Extension and application of credibility models in predicting claim frequency. *Mathematical Problems in Engineering*, *2018*, 1-8. https://doi.org/10.1155/2018/6250686

Yang, Y., Qian, W., & Zou, H. (2018). Insurance premium prediction via gradient tree-boosting Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, *36*(3), 456-470. https://doi.org/10.1080/07350015.2016.1200981

Yeung, J. S., & Wong, Y. D. (2015). Effects of driver age and experience in abrupt-onset hazards. *Accident Analysis and Prevention*, *78*, 110-117. https://doi.org/10.1016/j.app.2015.02.024

Yeung, K. (2016). Hypernudge: Big data as a mode of regulation by design. *Information, Communication & Society*, *20*(1), 118-136. https://doi.org/10.1080/1369118X.2016.1186713

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

Yip, K. C. H., & Yau, K. K. W. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance, Mathematics & Economics*, *36*(2), 153-163. https://doi.org/10.1016/j.insmatheco.2004.11.002

You, Y., & Li, X. (2017). Most unfavorable deductibles and coverage limits for multiple random risk with Archimedean copulas. *Annals of Operations Research*, *259*(1-2), 485-501. https://doi.org/10.1007/s10479-017-2537-9

Zhu, X., Yuan, Y., Hu, X., Chiu, Y.-C., & Ma, Y.-L. (2017). A Bayesian network model for contextual versus non-contextual driving behavior assessment. *Transportation Research*, *81*, 172-187. https://doi.org/10.1016/j.trc.2017.05.015

Zhuang, K., Wu, S., & Gao, X. (2018). Auto insurance business analytics approach for customer segmentation using multiple mixed-type data clustering algorithms. *Tehnicki Vjesnik*, *25*(6), 1783-1791. https://doi.org/10.17559/TV-20180720122815

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089. https://doi.org/10.1007/s10618-017-0506-1

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75-89. https://doi.org/10.1057jit.2015.5

Appendix A: Poisson Log Linear Model 4 with Coefficients

Table 10

Log Linear Poisson Model 4 with Coefficients

Poisson Log Linear Model with Coefficients

			Coefficient	Description	Wald Chi-Square	df	Pr > Chi Square
Target_Flag =	-2.208		(Intercept)		395.499	1	0.000
-	0.638	*	OFFENSE_CODE_0	Criminal Background Group 0	337.438	1	0.000
-	0.966	*	OFFENSE_CODE_1	Criminal Background Group 1	571.920	1	0.000
-	0.734	*	OFFENSE_CODE_2	Criminal Background Group 2	350.146	1	0.000
-	0.415	*	OFFENSE_CODE_3	Criminal Background Group 3	80.134	1	0.000
-	0.438	*	OFFENSE_CODE_4	Criminal Background Group 4	104.435	1	0.000
	0.110	*	OFFENSE_CODE_5	Criminal Background Group 5	1.699	1	0.192
-	0.323	*	OFFENSE_CODE_6	Criminal Background Group 6	49.098	1	0.000
-	0.137	*	OFFENSE_CODE_8	Criminal Background Group 7	5.655	1	0.017
-	0.210	*	OFFENSE_CODE_9	Criminal Background Group 94	8.609	1	0.003
	0.335	*	SALVEH	Compromised Vehicle	54.033	1	0.000
	0.418	*	RECAFA1	Recent At-Fault Accident 1	368.941	1	0.000
	0.253	*	RECAFA2	Recent At-Fault Accident 1	192.376	1	0.000
	0.149	*	RECMIN1	Recent Minimum 1	18.097	1	0.000
	0.171	*	RECMIN2	Recent Minimum 2	66.549	1	0.000
-	0.011	*	MINAGEINT1	Minimum Age Intercept 1	0.078	1	0.780
-	0.092	*	MINAGEINT2	Minimum Age Intercept 2	21.409	1	0.000
	0.043	*	MINAGEINT3	Minimum Age Intercept 3	0.991	1	0.320
	0.038	*	MINAGEINT4	Minimum Age Intercept 4	0.617	1	0.432
-	0.023	*	MINAGEINT5	Minimum Age Intercept 5	0.264	1	0.608
	0.112	*	SR22	SR 22 Filing	34.246	1	0.000
-	0.248	*	PRIORGRP31	Prior Insurance Group 1	223.377	1	0.000
-	0.031	*	PRIORGRP32	Prior Insurance Group 2	3.082	1	0.079
-	0.014	*	PRIORGRP33	Prior Insurance Group 3	0.582	1	0.446
	0.075	*	INCAGEINT1	Incurred Loss Age Intercept	6.321	1	0.012
-	0.530	*	HHVAR1	Household 1	137.772	1	0.000
-	-0.723	*	HHVAR2	Household 2	596.159	1	0.000
-	-0.467	*	HHVAR3	Household 3	76.547	1	0.000
-	-0.625	*	HHVAR4	Household 4	340.957	1	0.000
-	-0.405	*	HHVAR5	Household 5	21.381	1	0.000
-	-0.599	*	HHVAR6	Household 6	68.761	1	0.000
-	-0.409	*	HHVAR7	Household 7	7.714	1	0.005
-	-0.599	*	HHVAR8	Household 8	275.994	1	0.000
-	-0.338	*	HHVAR9	Household 9	60.438	1	0.000
-	-0.400	*	HHVAR10	Household 10	186.253	1	0.000
-	-0.137	*	HHVAR11	Household 11	7.471	1	0.006
-	-0.212	*	HHVAR12	Household 12	19.236	1	0.000
-	-1.366	*	HHVAR13	Household 13	1.864	1	0.172
-	-0.491	*	HHVAR14	Household 14	65.443	1	0.000
-	-0.321	*	HHVAR15	Household 15	12.134	1	0.000
-	-0.316	*	HHVAR16	Household 16	81.588	1	0.000
	0.482	*	CREDGRP1	Insurance Credit Score Group 1	89.729	1	0.000
	0.393	*	CREDGRP2	Insurance Credit Score Group 2	92.695	1	0.000
	0.259	*	CREDGRP3	Insurance Credit Score Group 3	62.068	1	0.000
	0.310	*	CREDGRP4	Insurance Credit Score Group 4	114.909	1	0.000
	0.233	*	CREDGRP5	Insurance Credit Score Group 5	49.614	1	0.000
	0.291	*	CREDGRP6	Insurance Credit Score Group 6	96.133	1	0.000
	0.299	*	CREDGRP7	Insurance Credit Score Group 7	103.887	1	0.000
	0.294	*	CREDGRP8	Insurance Credit Score Group 8	106.499	1	0.000
	0.203	*	CREDGRP9	Insurance Credit Score Group 9	54.438	1	0.000
	0.207	*	CREDGRP10	Insurance Credit Score Group 10	71.552	1	0.000
	0.116	*	CREDGRP11	Insurance Credit Score Group 11	23.820	1	0.000
	0.005	*	CREDGRP12	Insurance Credit Score Group 12	0.018	1	0.894
-	-0.168	*	CREDGRP13	Insurance Credit Score Group 13	18.148	1	0.000
	0.088	*	ATT1PNIINT1	At-Fault Primary NI Intercept 1	2.184	1	0.139
-	-0.046	*	ATT1PNIINT2	At-Fault Primary NI Intercept 2	0.321	1	0.571
	0.045	*	ATT1PNIINT3	At-Fault Primary NI Intercept 3	0.187	1	0.665
	0.069	*	ATT1PNIINT4	At-Fault Primary NI Intercept 4	1.106	1	0.293
	0.120	*	ATT1PNIINT5	At-Fault Primary NI Intercept 5	3.110	1	0.078
	0.167	*	ATT1PNIINT6	At-Fault Primary NI Intercept 6	5.615	1	0.018
	0.029	*	ATT1PNIINT7	At-Fault Primary NI Intercept 7	0.152	1	0.696
	0.197	*	ATT1PNIINT8	At-Fault Primary NI Intercept 8	8.827	1	0.003
	0.118	*	TERRGRP1	Territory Group 1	2.588	1	0.108
-	-0.228	*	TERRGRP2	Territory Group 2	5.880	1	0.015
	0.172	*	TERRGRP3	Territory Group 3	8.116	1	0.004
	0.128	*	TERRGRP4	Territory Group 4	4.294	1	0.038
	0.119	*	TERRGRP5	Territory Group 5	3.676	1	0.055
	0.229	*	TERRGRP6	Territory Group 6	12.341	1	0.000
	0.258	*	TERRGRP7	Territory Group 7	17.798	1	0.000
	0.174	*	TERRGRP8	Territory Group 8	7.418	1	0.006
	0.236	*	TERRGRP9	Territory Group 9	14.105	1	0.000
	0.238	*	TERRGRP10	Territory Group 10	13.745	1	0.000
	0.231	*	TERRGRP11	Territory Group 11	9.087	1	0.003
	0.206	*	TERRGRP12	Territory Group 12	10.533	1	0.001
-	-0.486	*	TERRGRP13	Territory Group 13	13.917	1	0.000
	0.260	*	TERRGRP14	Territory Group 14	10.759	1	0.001
	0.017	*	TERRGRP15	Territory Group 15	0.077	1	0.782
-	-0.046	*	TERRGRP16	Territory Group 16	0.439	1	0.507
	0.098	*	TERRGRP17	Territory Group 17	2.328	1	0.127
-	-0.013	*	TERRGRP18	Territory Group 18	0.043	1	0.835
	0.094	*	TERRGRP19	Territory Group 19	2.121	1	0.145
	0.047	*	TERRGRP20	Territory Group 20	0.571	1	0.450
	0.083	*	TERRGRP21	Territory Group 21	1.619	1	0.203
	0.067	*	TERRGRP22	Territory Group 22	0.321	1	0.571
	0.238	*	TERRGRP23	Territory Group 23	15.196	1	0.000
-	-0.056	*	TERRGRP24	Territory Group 24	0.472	1	0.492
-	-0.058	*	TERRGRP25	Territory Group 25	0.625	1	0.429
	0.207	*	TERRGRP26	Territory Group 26	5.630	1	0.018
-	-0.204	*	NSF3LATE1INT_1	NSF Payment Late Payment 1	27.273	1	0.000
-	-0.244	*	NSF3LATE1INT_2	NSF Payment Late Payment 2	39.340	1	0.000
-	-0.164	*	NSF3LATE1INT_3	NSF Payment Late Payment 3	16.823	1	0.000
-	-0.147	*	NSF3LATE1INT_4	NSF Payment Late Payment 4	10.058	1	0.002