COMPARISON OF CLASSIFICATION ALGORITHMS AND UNDERSAMPLING METHODS ON EMPLOYEE CHURN PREDICTION: A CASE STUDY OF A TECH COMPANY

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Industrial Engineering

by

Heather Cooper

December 2020

© 2020

Heather Cooper

ALL RIGHTS RESERVED

COMMITTEE MEMBERSHIP

TITLE: Comparison of Classification Algorithms and Undersampling Methods on Employee Churn Prediction: A Case Study of a Tech Company

AUTHOR: Heather Cooper

DATE SUBMITTED: December 2020

COMMITTEE CHAIR: Tali Freed, Ph.D.
Professor of Industrial Engineering
California Polytechnic State University

COMMITTEE MEMBER: Roy Jafari, Ph.D.
Assistant Professor of Business Analytics
University of Redlands

COMMITTEE MEMBER: Kurt Colvin, Ph.D.
Professor of Industrial Engineering
California Polytechnic State University

ABSTRACT

Comparison of Classification Algorithms and Undersampling Methods on Employee Churn Prediction: A Case Study of a Tech Company

Heather Cooper

Churn prediction is a common data mining problem that many companies face across industries. More commonly, customer churn has been studied extensively within the telecommunications industry where there is low customer retention due to high market competition. Similar to customer churn, employee churn is very costly to a company and by not deploying proper risk mitigation strategies, profits cannot be maximized, and valuable employees may leave the company. The cost to replace an employee is exponentially higher than finding a replacement, so it is in any company's best interest to prioritize employee retention.

This research combines machine learning techniques with undersampling in hopes of identifying employees at risk of churn so retention strategies can be implemented before it is too late. Four different classification algorithms are tested on a variety of undersampled datasets in order to find the most effective undersampling and classification method for predicting employee churn. Statistical analysis is conducted on the appropriate evaluation metrics to find the most significant methods.

The results of this study can be used by the company to target individuals at risk of churn so that risk mitigation strategies can be effective in retaining the valuable employees. Methods and results can be tested and applied across different industries and companies.

Keywords: Churn Prediction, Data Mining, Unbalanced Datasets, Machine Learning

ACKNOWLEDGMENTS

I'd like to first thank my parents for their overwhelming support throughout college and for providing me the privilege of pursuing higher education. I would also like to thank my close friends, Maggie and Joe, for showing me how cool data and statistics can be. I deeply value our friendship and the levity you bring to my life. I'd also like to extend thanks to my biggest supporters, Viraj and Jake, for their continued support and being my voices of reason.

Thank you to my advisor, Dr. Freed, for not only taking me on in my last quarter of research but for being a constant in my final years of college. Thank you to Dr. Colvin for the critical and constructive feedback on my research.

Lastly, I'd like to thank my mentor, Dr. Jafari, for suggesting this topic to me and assisting me every step of the way on this thesis. At the beginning of college, I would not have imagined I'd be where I am today with my education and career goals, and I attribute this change to your courses and passion for learning.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

1. INTRODUCTION

1.1 RESEARCH MOTIVATION

The focus on data driven decision making has increased as business processes are being changed to align with what the data conveys rather than based on employee intuition and observation alone. Predictive modeling can be an asset to a company as it shows patterns and trends in the data that may not be recognized by a human operator. Data modeling and analysis can be utilized to improve business processes and company standards to achieve a predefined goal. The biggest advantage in using data driven decision making is that the algorithm and analytics can recognize patterns in the dataset that could otherwise be overlooked by an analyst. Results will remain consistent over time rather than being clouded by human judgement.

Churn prediction is a well-researched area of advanced analytics within any industry. Churn can be thought of as the annual rate that employees or customers leave a company. The focus of this research is to predict an employee's churn from a telecommunications company in time to apply risk mitigation strategies to prevent the employee from deciding to leave the company. Finding the best fit for a position in an organization can be a costly and lengthy process; any company needs to have a system in place to lower their employee turnover rate. By lowering the employee churn rate, the company ensures that they must go through the hiring process as infrequent as possible. Increasing retention is in the best interest of any business as it reduces overhead hiring cost, saves time, and better allocates company resources.

Within high tech industries, companies are facing a higher-than-average employee churn rate of 12-15% (Alamsyah & Salma, 2018). This could be due to the exponentially

growing opportunities for people in these technical fields as businesses continue to grow and new jobs are created. If an employee from a technical position were to churn, it would be harder to find a replacement for the position as the qualified applicant pool would be more limited. In addition, if a company has a high turnover rate, prospective applicants would be turned off from applying there as it implies that they are unable to keep their employees happy. This would further increase the hiring and recruitment cost until the high churn rate is investigated and lowered.

Beyond these explicit savings, the amount of training and resources invested into a new hire must also be considered when discussing the significance of reducing employee churn. For the first period of a new hire's role, they offer little to no profit to the company as they are continuously job shadowing, training, or learning the company's business practices and the requirements of their position. This is a worthy investment the company initially makes in the new employee as they recognize that the years of potential work that employee has will more than pay off the initial deficit. If an employee churns, especially within the first two years of starting their position, the organization likely won't see a profit from their work. They may even lose money on their initial investment of hiring the employee as they left so prematurely.

1.2 RESEARCH SIGNIFICANCE

It is important to determine the factors and attributes that increase the likelihood of an employee to churn and adopt risk mitigation protocols in response to higher risks. This will help management and human resources identify and try to reach out to the high-risk employee before there is no chance of retaining them, leading to savings in hiring costs

and time. By being able to pinpoint common reasons for churn and focus on these attributes, the company can reassess and redefine their common business practices to better suit what their employees need. Changing the company's standard procedures to align with what their employees need has the potential to increase general job satisfaction across the board. To do this, machine learning algorithms will be applied to past data from an Iranian telecommunications company in hopes of creating a model that will successfully classify and rank future employees based off their churn risk as well as identify attributes that are key contributors to churn.

1.3 RESEARCH QUESTION

Can undersampling and machine learning techniques be applied to employee data in order to properly identify those at risk of leaving their company? Further, can decision rules and heuristics be pulled from these models to provide insight into the reason for employee churn, leading to changes in business processes and practices to increase a company's retention rate?

2. BACKGROUND

2.1 CLASSIFICATION MODELS

Within data mining, there are three different kinds of models that can be used to solve an enterprise problem. These include prediction, classification, and clustering models. Prediction models will return an estimated (or predicted) value for data objects based off the historical data. Clustering models are often used as a preprocessing step; these models will subset the data by similar attributes or behaviors amongst the data objects, indicative of a certain class. For churn prediction, the model used is classification, as there are only two outcomes for the data objects: churn and non-churn. Classification algorithms are the most widely used in data mining, as they are tailored for supervised learning where outcomes of the data are already specified and defined (Dogan & Tanrikulu, 2013). A limitation of classification models is that they will find patterns and trends in the data that lead to the specified classifications, but there might be a separate outcome that can better be predicted by the dataset.

There are many different algorithms that can be used to solve classification data mining problems. The algorithms most used in literature to predict churn are Naïve Bayesian, Decision Tree, Artificial Neural Networks, and Support Vector Machines.

2.1.1 Naïve Bayesian

Naïve Bayesian classification applies Bayes' Theorem and utilizes probability and assumption of event independence to calculate scores for each data object belonging to a certain class. These scores are probabilistic-like values where a higher score implies a greater likelihood of a data object belonging to the specified class, but they differ in that

the score for an object belonging to a class and not belonging to a class don't have to have a sum of one. Naïve Bayesian requires numeric inputs; categorical or non-numeric data can be transformed through discretization or by assuming a probability distribution to prepare data for classification. This model relies heavily on independence between attributes, meaning the value of one independent attribute has no correlation or effect on other independent variables. However, this is rarely true in real world applications; for instance, employee tenure and age can be correlated since an employee's tenure is limited by their working age. Even though this leads to a lower prediction performance, the algorithm still assumes independence which keeps model complexity low and interpretability high (Jafari, 2020).

2.1.2 Support Vector Machines

Support Vector Machine (SVM) uses similar concepts as regression analysis, except the algorithm has a higher performance while complexity and computing costs are still relatively low. The model uses the dataset of "N" dimensions, where "N" is the number of independent variables in the dataset, to identify a hyperplane of N-dimensions to classify the data objects. This hyperplane is identified by finding the maximum distance between both binary classes in order to reduce error and uncertainty with future data points. Once the plane is identified, the highly influential points, known as the support vectors, are the data points located near the hyperplane and the margin surrounding the plane. Adding or taking away support vectors will greatly affect the coordinates of the hyperplane, shown in Figure 1; this is what helps build a successful model. The data

point positions in relation to the hyperplane is calculated into a score, and this score is then interpreted into a classification into one of the binary classes. (Fletcher, n.d.)

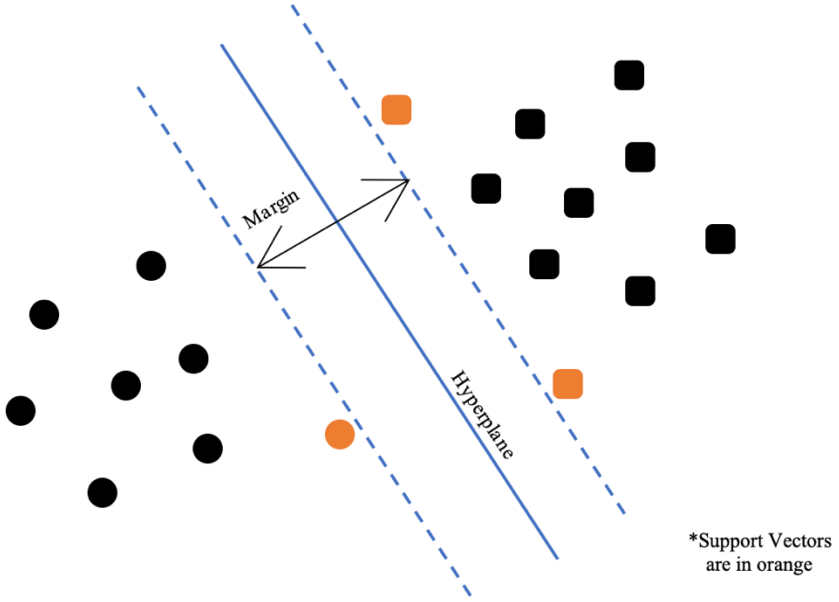

Figure 1: Example of an SVM

SVM is advantageous for its increased classification power while remaining relatively interpretable. However, as problem complexity increases and attribute dimensions grow, the results become less intuitive. A hyperplane that's for a two- or three-dimensional dataset would be a line or a plane; any dataset of greater dimension wouldn't be as easy to understand. Even though its transparency decreases with more attributes, SVM has proven in literature to be a successful classifier in churn prediction.

2.1.3 Decision Tree

Decision Trees are sets of decision rules created from a specific dataset in order to classify the given data objects. The final output of a decision tree is a tree-shaped structure that acts as a flowchart for data objects to be fed through and classified based on

their attribute values. Decision trees are comprised of split and leaf nodes. Split nodes contain an attribute test for the data object to either move onto another split node or a classification. Leaf nodes are the ends of the split nodes that classify the data object, depending on the enterprise problem. The objective is for each split node to further divide the dataset into purer subsets based on the dependent variable (Jafari, 2020). Ideally, when a dataset passes through a decision tree the dataset splits so that at each leaf node there's a subset of the data all belonging to one class. Figure 2 shows an example of a simple decision tree that could be used to classify customer churn.

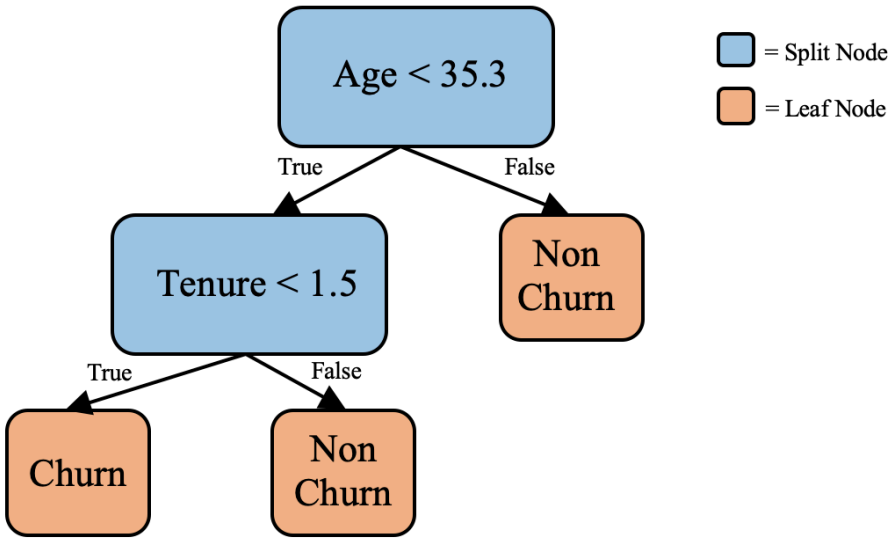

Figure 2: Example of a Decision Tree Classifying Churn

Decision Tree algorithms require extensive tuning to avoid overfitting the data to the model. This includes finding an optimal depth (levels of split nodes), sample split, and impurity decrease through testing. An obvious advantage to using decision tree for classification is its extremely easy to interpret and extract information on what common attributes are for each class. However, this model is not as capable at handling complex non-linear relationships between attributes as other algorithms discussed in the next

sections. Decision tree has shown significant success in predicting churn, depending on the dataset and attribute types.

2.1.4 Artificial Neural Networks

Artificial Neural Networks (ANN) are designed to simulate human learning by using a complex innerweb of neurons and weighted connections in order to predict or classify data objects. ANN is especially successful with complex, non-linear relationships between independent and dependent variables. The most common ANN algorithm is Multilayer Perceptron (MLP), or feedforward ANN. In an MLP structure, each independent variable in the dataset will have an input neuron and every dependent variable will have an output neuron. Between these there's a complex web-like structure of hidden layers and neurons with weighted connections that connect the data object from the given inputs to the predicted output (Jafari, 2020). Figure 3 shows an example of an MLP structure for a churn prediction model. Initially, a random weight is assigned to each connection and data objects are introduced to the untrained model. As data objects are fed through the model, the weights are adjusted as the model learns from its classification of each data object. The model is considered fully tuned when there is not significant change in weights after feeding the dataset through. The more complex and non-linear a problem, the more hidden layers and neurons are needed. With more hidden layers, the user risks overfitting the data and increasing computational costs; there needs to be a balance of fitting to complexity and minimizing computational costs.

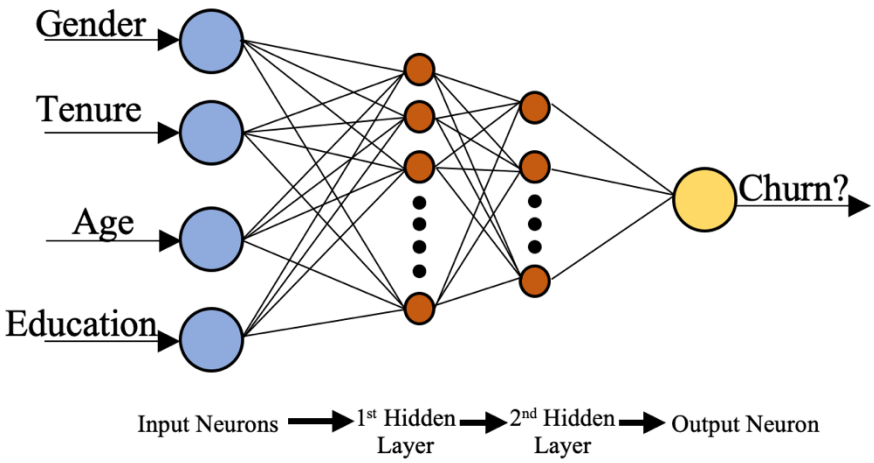

Figure 3: Example of a MLP Structure in Classifying Churn

An advantage in using ANN is its higher accuracy, resiliency towards outliers, and success with more complex problems. Since the algorithm is inspired by the complex way that the human mind works, it is usually able to achieve a higher success with the enterprise problem than other machine learning algorithms. However, the results of ANN or MLP models are not as interpretable, so it is harder to extract useful information about the problem at hand. MLP is often referred to as "black box" classification, as the steps that happen between input and output are not very clear to the user.

2.2 CLUSTERING

As mentioned, clustering is another method of solving enterprise problems with machine learning. While prediction and classification algorithms provide a measurable value for each data object, clustering finds unique patterns based on data attributes and groups them based off these similar features. The groups don't always have a correlation with the dependent variable, but they may provide further insight to how the company can use prediction and classification within the cluster groups (Jafari, 2020).

In research, clustering has been used as a preprocessing step to other data mining tasks, such as churn prediction, to increase their performance. Cluster groups can be added to the original so the clusters can be used as an independent variable in classification or prediction. They can also be used to divide the original dataset into cluster subsets, and then apply data mining algorithms to the subsets. Dividing the original dataset into the clusters ensures that the classification and prediction algorithms identify more unique patterns specific to the clusters, which improves the overall performance.

2.3 COST-SENSITIVE CLASSIFICATION

Cost-Sensitive Learning considers different costs associated with misclassification of the classes within the data preprocessing or model tuning with the goal of minimizing total cost. For binary classification, it is usually assumed that a misclassification of either class, a false negative or false positive, has an equally negative impact on the company; this is often not the case, especially in unbalanced datasets where identifying members of the minority class is the company's priority. Think of medical exams used to test for cancer in patients: a false positive, or Type I error, means a patient was identified as at risk of cancer when they don't actually have any higher risk than the average person. False negatives, or Type II error, are cases where patients are identified as non-risk when in fact, they do have cancer or are at higher risk of developing cancer. Type II errors have a higher associated cost, as being told you don't have cancer when you do can lead to delayed treatment or medical attention, which could ultimately lead to death; Type I errors are likely to be corrected with further testing before any procedures or treatments are started. Therefore, during classification it is more important to prioritize identifying

the positive cases than the negative, even if that consequently leads to an increase in the number of false positives.

There are two ways to address the cost-sensitivity of classification: the direct or cost-sensitive meta learning method. The direct method involved designing the classifiers to be cost sensitive and takes place during the model building. Cost-sensitive meta learning includes adjusting the threshold for classification based on the dataset, under or oversampling, and weighting the data objects based off their misclassification cost. By accounting for the difference in misclassification costs for churn prediction, the algorithm's success in correctly classifying the at-risk employees increase and there's a greater chance of retention.

2.4 UNBALANCED DATASETS

With classification churn problems, it is often the case that the class of interest makes up a much smaller proportion of the available data than what is assumed for the algorithms. For churn prediction, the number of people leaving a company is – hopefully – significantly smaller than the proportion that stays. This class imbalance violates the assumption of most classification algorithms that each class makes up an equal proportion of the dataset, so the performance and results are usually skewed.

The most common methods to address class imbalance are oversampling and undersampling. These take place in the data preprocessing stage and reduce the imbalance before the algorithms are applied to the data. Oversampling involves selecting data objects from the minority class and replicating them until the two classes are balanced. While this is the simplest method to address class imbalance, it also increases

the risk of overfitting the data. On the other hand, under sampling focuses on the majority class and identifies redundant data objects to delete from the dataset. However, with deleting data objects, there's a chance that crucial data objects won't be included in the model, weakening the final results. In literature, it is more common for under sampling to be implemented rather than oversampling when classifying churn. Sundarkumar et al used under sampling to address class imbalance in their dataset (Sundarkumar & Ravi, 2015). By using k Random Nearest Neighbors method to identify and eliminate outliers from the original dataset, they were able to identify non-significant data objects that were least likely to be influential in classification. After applying kRNN, one class SVM was used to further under sample the majority class, resulting in a nearly perfectly balanced dataset that performed better with classification than when these methods were not used. This leads to the conclusion of the importance of applying random-based models to the data coupled with under or over sampling in order to avoid losing valuable information for classification.

2.5 RECURSIVE FEATURE ELIMINATION

Recursive feature elimination (RFE) is a method used in data preprocessing to select the significant attributes to retain in data mining models. The inputs required for RFE are the number of desired features, the evaluation metric used to assess performance, and the classifier used for each trial. The algorithm start factoring all variables into the analysis and one by one will eliminate an attribute from the usable pool. At each level, all attributes are temporarily removed and the evaluation metric for the classifier is recorded;

the attribute that least effects the evaluation metric will be dropped. This process is repeated until the specified number of features is met.

RFE is a valuable process for data preprocessing for dimension reduction and addressing data redundancy. Including all attributes for analysis does not necessarily lead to a better classification performance; rather, the adverse effect could occur, and performance could be hindered from excess, overlapping information. Finding the balance between enough features to capture as much information about the data objects as possible while not overfitting the model to the specific data is challenging, yet pivotal for the success of the classification algorithm.

2.6 MODEL EVALUATION

There are two steps to evaluate success of data mining models: during model construction and after results are generated. During model definition, in order to avoid overfitting, the dataset is usually split into training and testing subsets. The training set is the data used to construct and tune the data mining model, while the testing subset is used post-construction to verify and validate the results of the model. The training dataset is usually around 70-85% of the original dataset, with the remaining belonging to the testing set. To further validate the model's performance, a sampling method known as k-Fold Cross Validation can be applied, shown in Figure 4. K-Fold Cross Validation splits the entire dataset into k groups; k algorithms are then tuned with each having a different group assigned as its test set.

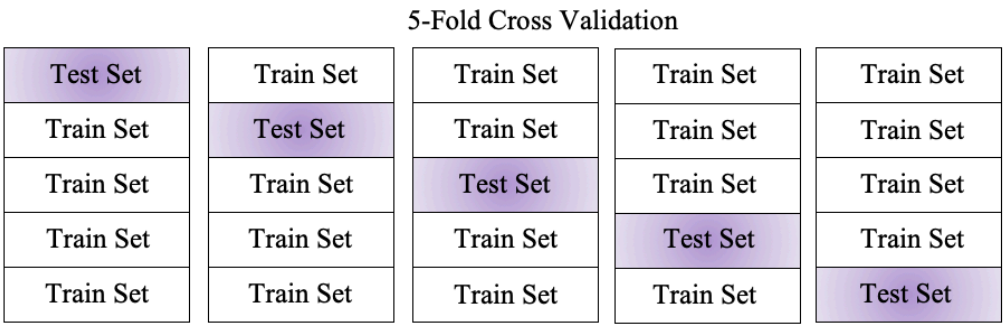

Figure 4: Example of k-Fold Cross Validation, with k = 5

Once all algorithms' performance measures are recorded, the average and variance of the performance metric is used to evaluate the success of the models. The model with the best performance is then retained for future use.

For classification data mining problems, once results are obtained, a quick way to interpret the output is through a confusion matrix. A confusion matrix (see Figure 5) shows the frequency of true positive, false positive, true negative, and false negative classifications from the model in a table. This is helpful later in calculating metrics of performance.

		Predicted: Non-Churn	Predicted: Churn
Actual	Non-Churn	TN	FP
	Churn	FN	TP

Figure 5: Example of a Confusion Matrix for Churn Classification

From the confusion matrix, there are four common evaluation metrics used in literature and these are accuracy, precision, recall, and F-measure. Accuracy is the proportion of data objects that were correctly classified compared to the number of data objects in the entire set. For imbalanced datasets, accuracy is not a good measure to use to evaluate a model's performance since the dataset is highly imbalanced. A high accuracy may seem successful at first, but after accounting for the large proportion of data that is a part of the non-churn class it is easy to be misled by an algorithm that unsuccessfully classifies every data object as non-churn.

Another shortcoming of using accuracy as an evaluation metric is it evenly weights both false negatives and false positives. In enterprise problems, these misclassifications likely have different levels of importance and thus need to be prioritized accordingly. A higher recall score is ideal when false negatives need to be minimized, while a high precision score will lead to a reduction of false positives (see Figure 6).

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Figure 6: Formulas for Recall and Precision

The F-measure is a balance of recall and precision scores, evaluating the classifier performance as a whole (Figure 7). The closer to one, the better the balance between recall and precision for the classification algorithm.

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 7: Formula for F-Measure

As mentioned, the greatest drawback to using these metrics is that they have a heavy bias when used to evaluate unbalanced datasets. There are other evaluation metrics that are used in cases where the data object sameness assumption cannot be made, the most common being the receiver operating characteristic (ROC), shown in Figure 8, and area under the ROC curve. ROC is used for binary classification evaluations, plotting the true positive rate against the false positive rate as threshold values change. Threshold values are only applicable with algorithms that output probabilistic like values, such as Naïve Bayesian or MLP.

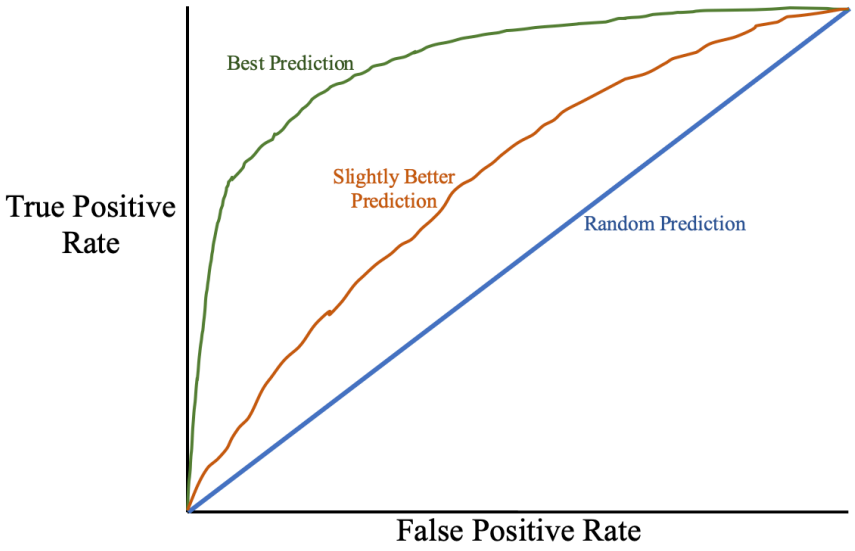

Figure 8: Example of ROC

When interpreting an ROC curve, the straight diagonal line bisecting the middle of the graph represents the random classification performance. The more exaggerated the curve above the random line, the better the classification performance. A threshold value is selected based on the elbow point of the curve, which would maximize true positive rate while minimizing the false positive rate.

Another metric that builds off the ROC curve is the area under the ROC curve (AUC-ROC), see Figure 9. AUC-ROC is a summary of the ROC and measures the classifiers ability to classify the data objects. It is calculated by finding the area under the ROC curve; the closer to one, the better the classifier. If the AUC-ROC is less than 0.5, the classifier isn't performing any better than a random guess.

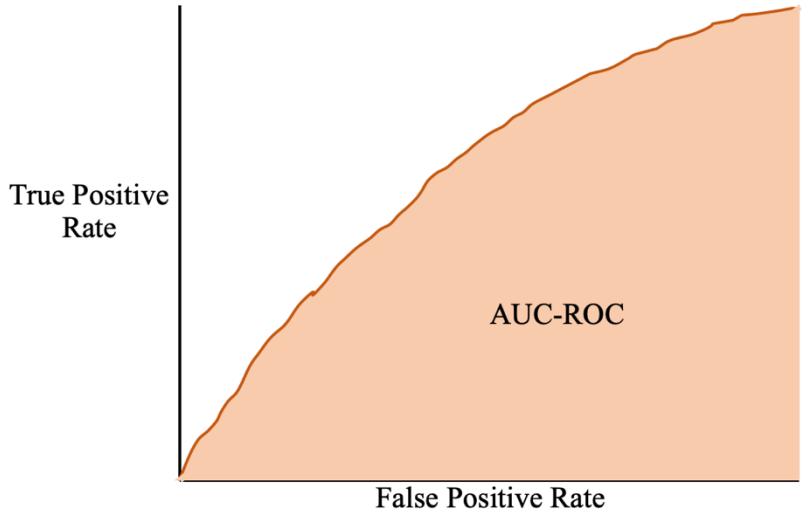

Figure 9: Example of AUC-ROC

3. LITERATURE REVIEW

3.1 INTRODUCTION TO LITERATURE REVIEW

The purpose of this literature review is to give the reader an overview of different machine learning and data mining methods and how they have been applied in industry churn prediction. More advanced analytics such as predictive and classification modeling has been reserved for larger companies and corporations that have the proper infrastructure in place to produce good data to be used in the algorithms. As data has become more accessible and utilized, a greater number of middle market companies have started to use advanced analytics within their practices to help preserve their business.

3.2 CUSTOMER CHURN

There is already a vast amount of research focusing on customer churn prediction, specifically within the telecommunication industry where competition is high and customer loyalty is quickly changing. The cost associated with recruiting new customers far exceeds retaining current ones; therefore, it is in any company's best interest to invest in ways to predict the customers that are at risk of churning so they can focus their efforts on keeping them with the company. Studies have shown that the key difference between customer and employee churn is the reasoning behind the individual leaving the company, as this greatly depends on the relationship between the churner and the company. Customers will churn for many non-complex reasons, such as an influential leader in their social group leaving their provider for a better service. If a leader in the customer's circle leaves their provider, the churn risk of that entire social circle increases (Richter et al., 2010). These reasons are easier to portray with the correct data; research

has shown success in using rule-based decision making in creating decision rules for customer churn (Tsai & Chen, 2010).

The most important step in churn prediction is preparing the data appropriately for the algorithms selected. There are many different techniques that have been studied to determine what will result in the best classifiers for a model. In Vafeiadis et al., boosting was used to iterate over the weak classifiers in the dataset; this assigned new, different and higher weights to the misclassified objects which were deemed significant cases in successful churn prediction. By putting a higher weight on learning from the misclassified data objects, the resultant dataset is a set of "highly accurate" classifiers that can be used on the validation data set for testing their success in churn prediction (Vafeiadis et al., 2015).

Tsai and Chen also focused their research on effective data preprocessing in order to improve churn classification. In their paper, they focused on sorting out data objects from the original dataset that did not represent the population of interest in order to properly train the classification algorithm on more typical data objects. They also used association rules in attribute selection to narrow down the characteristics in the model and include ones that significantly contributed to churn. An importance score was calculated for each attribute and only ones with a score greater than 0.8 (on a 0-1 scale) were selected to be included in the model. The combination of association rules being used in both under sampling and feature selection led to a significant improvement in classification performance compared to when neither were used (Tsai & Chen, 2010).

After identifying the churn risks within your dataset, a challenge that most companies face is being able to act and lower the churn risk for their customers before it is too late.

It is not feasible to try and recover all the customers classified as churn as that will likely prove to be unsuccessful and fail to recover the company's lost profits. Lazarov and Capota factored in a customer's lifetime value based on a few key attributes including cash flow, loyalty to the company, and effect within their social group, to prioritize customers that were identified as churn risks and only target those that also had a high lifetime value (Lazarov et al., 2007). Focusing retention efforts on customers with a high lifetime value with the company allowed only the most profitable customers to be targeted for retention while the less profitable ones were not given as high of a priority. This will better ensure that the profits due to retention efforts are maximized and the more valuable customers are retained, when possible.

In order to be able to apply effective retention efforts, it is necessary to identify the reasoning behind a company's churning customers. The algorithm used to classify churn plays an important role in the interpretability of the results for the company. If artificial neural networks or support vector machines are implemented, the results won't be as interpretable for the company to develop new business processes as the algorithms work within a black box to classify data objects. Decision Tree, Naïve Bayesian, or Logistic Regression all have more interpretable outputs that can be used to help modify business practices. Amin et al utilized rough set theory to create decision rules structured as "IF - THEN" statements that guide the classification of customers. These rules provide great insight into the contributing factors for customers leaving, but the overall performance and accuracy of the algorithm compared to the black box methods is not as high (Amin et al., 2017).

3.3 EMPLOYEE CHURN

Employee churn uses past data to identify and predict the employees that are most likely to leave the company based off of a given set of attributes. There are two types of churn for employees: involuntary and voluntary. Involuntary churn is when an employee exits a company for reasons beyond their immediate control or decisions. This could include layoffs, firing, or transfers to different departments. Voluntary churn is the focus of employee churn prediction, as it includes employees who leave a company on their own accord for various reasons. These reasons can be positive, such as better opportunity, pay, or incentives elsewhere, or for negative reasons like lack of interest, no growth, or poor working conditions. Because employee churn is often due to more personal reasons and preferences than customer churn, it is especially challenging to procure data that can accurately predict all churn risks. For these classification models, the data is generally from a company's Human Resources Information System (HRIS), which is saturated with noise due to the method of collecting the information and validity of the data after time. Employee churn datasets also have a highly imbalanced dataset, with a small percentage of data objects classified as churners compared to employees that stayed with the company. Addressing the imbalance in the data as well as the reliability of the data source are important preprocessing steps before creating employee churn models to ensure the results are valid and can be trusted.

When an employee leaves a company unexpectedly, the company is adversely impacted in many ways. Their departure can result in their workload being distributed, creating an unfair balance on remaining employees. This imbalance can lead to a time and money loss as project deadlines may need to be extended to accommodate the increased

workload, which could lead to a higher customer dissatisfaction. Attrition also has a negative effect on company morale and in certain settings may spark a chain reaction of employees leaving. Overall, it is more expensive from a company's standpoint to replace an employee than a customer due to the smaller pool of possible applicants for an employee's position, as well as the indirect costs detailed above. Replacing an employee is especially costly in high tech industries since there are many job opportunities available to an even smaller subset of people. Without focusing on employee retention and lowering attrition, then a company cannot truly reach their maximum profit potentials.

Feature selection in churn prediction provides insight into the common trends behind high attrition rates. Ma et al found that an employee's relationship with their supervisor and job satisfaction were good indicators of their likelihood to churn. In this study, researchers implemented an app with several different companies to collect satisfaction ratings and feedback daily from the employees. This data was then used to predict their likelihood of churn based on their feedback in comparison to the company's averages (Ma et al., 2019). Other attributes that had a high influence on churn was gender and ethnicity. Alao et al found that at different ages, gender had a different correlation with other employee attributes depending on the stage of life the person was in and the gender's expectations of that time. For example, at the time of the study in the country of origin it was expected that women would stay home and care for their children; the research found a strong positive correlation between employee churn in females with multiple children. Ethnicity also proved to be a significant contributor in churn prediction

– some cultures place a higher value on loyalty, which can translate to how acceptable an employee would find leaving their company (D & B, 2013).

With employee churn, it's important to understand that every employee that is classified as churn risk isn't necessarily worth the extensive efforts to retain. It should be a company's priority to identify and retain the employees that are a churn risk who bring the most value and competitive edge to the company. In Saradhi and Palshikar's research, they calculated an employee's value based on their projects, billable months, and on/off site assignments and used this measure to prioritize their list of high-risk churners. Since employees will churn for more complex reasons than customers, strategies to retain employees are also more complicated and specialized based on the individual. It is nearly impossible to implement one all-encompassing method of retention that will be equally valued and beneficial to every churn risk. By identifying the most important employees to retain, the company can properly address and manage their efforts so that the likelihood of retention for the most impactful employees is improved (Saradhi & Palshikar, 2011).

In a similar study, Shankar et al also applied employee value to prioritize risk mitigation efforts by using metrics such as an employee's environmental satisfaction, performance rating, and job level to calculate value (Shankar et al., 2018).

There has been extensive research into the best algorithm to use in employee churn prediction, what features are most significant, and which evaluation metrics provide the best interpretation of success. Table 1 below is a summary of all the sources that were consulted for this project and important takeaways from each piece of research. The bolded methods were the most successful churn predictors for that piece of research, and the attributes with an asterisk were found to be most significant for the study.

Table 1: Summary of Employee Churn Literature Review

Source	Summary	Industry	Methods Used	Evaluation Metrics	Attributes Used
(Hong & Chao, 2007)	Applying classification algorithms probit and logit to use a non-linearly combined variables on an employee churn dataset	Motor Market	Probit and logit models	F-Measure, AUC-ROC, True Positive Rate, False Positive Rate, Precision	Sex, DOB, State of Origin, Grade Level/Step, Tenure*, Salary*, Reason for Leaving
(Alamsyah & Salma, 2018)	Compare performance of three common classification algorithms to find the best performer, using 10-fold cross validation to further validate performance	Telecom	Naïve Bayes, Decision Tree, **Random Forest**	True Positive, False Positive, True Negative, False Negative	Location, Division, Sub-Division, Position, Level of Employee Within Organization, Age, Sex, Tenure, Salary, Take Home Pay
(Chang, 2009)	Applied a new method of feature selection that combined the Taguchi method with Nearest Neighbor Selection Model to create data subsets and select the best factors for predicting employee churn	Not Specified	N/A	Accuracy	Sex, Age, Education Level, Marriage, Resident, Full/Part Time, Salary, Department, Position, Prior Experience, Tenure, Average Validated Credit, Average Age of Children, Partner's Employment, Working Hours, Company

					Involvement, Sick Leave
(D & B, 2013)	Generate decision tree models and rule sets from dataset in order to correctly classify churn	Higher Institution	Decision Trees (CART, REP, C4.5)	Quadratic Probability Score, True Positive, False Positive, Precision, F-Measure, AUC-ROC	Sex*, DOB, State of Origin, Grade Level, Tenure, Salary, Reason for Leaving, Ethnicity*
(Dutta & Bandyopadhyay, 2020)	Using 10-fold cross validation with feed forwarding neural network, compare the performance at classifying employee churn against other popular classification algorithms	Not Specified	SVM, KNN, Naïve Bayesian, Decision Tree, Random Forest, Adaboost Classifiers	Accuracy, Precision, Recall, F-Measure	Age, Salary, Distance from Home, Education, Environmental Satisfaction, Number of Companies Worked, Performance Rating, Relationship Satisfaction, Stock Market Option, Standard Hours, Tenure, Training Hours, Balance in Work Life, Years Since Last Promotion, Relationship with Current Manager in Years
(Jain, 2017)	Focuses on data mining and feature extraction in order to increase performance of classification algorithms	IBM	KNN, Generalized Linear Model, SVM, Random Forest, Decision Tree,	Accuracy, Sensitivity, Specificity	Age*, Marital Status*, Sex*, Job Satisfaction*, Monthly Salary*, Work Life Balance*, Ethnicity*, Education*, Job Position, Stock Option Level,

			Adaptive Boosting		Relationship Satisfaction, Overtime, Job Involvement
(Jantan et al., 2009)	In a very broad sense, covers applying data mining within HR and Talent Acquisition practices	Not Specified	**Random Forest,** MLP, Radial Basic Function Network	Accuracy	Not Specified
(Ma et al., 2019)	Using a publicly available dataset of employee feedback and satisfaction from various companies, apply churn models to determine from a subset of significant features employee churn and how it differs by company	Various Companies	**Decision Trees**, NB, MLP	AUC-ROC	Various self-reported measurements from an app about employee satisfaction and feedback
(Punnoose & Ajit, 2016)	Applying Extreme Gradient Boosting to the noisy HRIM data in order to better classify employee churn compared to other popular algorithms	Retail	Logistic Regression, Naïve Bayesian, Random Forest, KNN, Linear Discriminant Analysis, SVM, Extreme Gradient Boosting	AUC-ROC, algorithm run time and memory utilization	Age, Tenure, Salary, Job Satisfaction, Employee's Perception of Fairness, Sex, Ethnicity, Education, Marital Status
(Saradhi &	Beyond correctly classifying	Not Specified	**SVM**, Decision	Accuracy, True	Age, Sex, Department,

Palshikar, 2011)	employee churn, it is also important to factor in employee value to better prioritize actions after churn		Tree, Naïve Bayesian, Logistic Regressi on	Positive, True Negative	Qualification, Tenure, Location, Experience in Parent or Client Organization, Billed or Not, On/Off Site, Designation in Client Organization
(Shankar et al., 2018)	Using clustering methods within departments, researchers categorized employees as churners based on their score in performance, education, and salary and how it compared to their department's average. After classification, used job level, performance rating, and environmental satisfaction to target best employees who were at risk of churn.	IBM	Decision Tree, Logistic Regressi on, SVM, KNN, Random Forest, Naïve Bayesian	Not Specified	Age, Business Travel, Daily Rate, Education, Distance from Home, Department, Education, Sex, Environmental Satisfaction, Salary, Job Level, Job Involvement, Job Role, Marital Status, Job Satisfaction, Over Time, Number of Companies Worked, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Stock Option Level, Tenure, Training Time, Work Life Balance, Years Since Last Promotion
(Yigit & Shourabiz adeh, 2017)	Utilized recursive feature elimination to select features for	IBM	Decision Tree, Logistic Regressi	Accuracy, Precision, Recall, F-Measure	Education, Environmental Satisfaction, Sex, Job Involvement,

	the model and compared the performance of the different algorithms without this feature; all algorithms were more successful churn predictors when recursive feature elimination was used compared to not.		on, **SVM**, KNN, Random Forest, Naïve Bayesian		Job Level, Job Role, Job Satisfaction, Marital Status, Over-Time, Performance Rating, Relationship Satisfaction, Stock Option Level, Years Since Last Promotion, Years with Current Manager
(Zhao et al., 2019)	Splits two datasets into small, medium, and large subsets and compares the predictive performance of all algorithms on each dataset with three different data formats applied: raw, standardized, and normalized.	Bank	**Decision Tree, Random Forest, Gradient Boosting Tree**, Extreme Gradient Boosting, Logistic Regression, SVM, Neural Networks, Linear Discriminant Analysis, Naïve Bayesian, KNN	AUC-ROC, Accuracy, Precision, Recall, F1	Last Pay Raise*, Tenure, Age, Compensation, Specialized Area, Department, Education, Background, Performance, Team

4. DATA

4.1 INTRODUCTION

An Iranian telecommunications company been experiencing higher than expected employee churn rates and intends to investigate and pinpoint possible reasons for their employee's churn. The data has been collected over the course of several years on a quarterly basis. The goal for this study is to develop and adapt business practices in order to reduce employee churn between data assessment periods.

4.2 FEATURE SELECTION

4.2.1 From Literature

Feature extraction is one of the most important data preprocessing steps as it leads to more significant and efficient machine learning models. In literature, there were thirteen common attributes that were consistently significant predictors of employee churn across industries and data sets. The significant attributes, as well as their appearance in sources, can be found in Table 2. Looking through the list, most of these attributes would already be readily available in an HR database, such as age, tenure, gender, location, and education level. A few of the attributes require an assessment or survey of the employees, while others would need even further investigation. After interviewing with professionals familiar with churn prediction, all the attributes found to be significant in literature were also thought to be of importance from a logical standpoint.

Table 2: Most Common Attributes Used for Employee Churn Prediction

Attribute	Source
Age	(Alamsyah & Salma, 2018; Chang, 2009; D & B, 2013; Dutta & Bandyopadhyay, 2020; Hong & Chao, 2007; Jain, 2017; Saradhi & Palshikar, 2011; Zhao et al., 2019)
Tenure	(Alamsyah & Salma, 2018; Chang, 2009; D & B, 2013; Hong & Chao, 2007; Saradhi & Palshikar, 2011; Zhao et al., 2019)
Gender	(Alamsyah & Salma, 2018; Chang, 2009; D & B, 2013; Hong & Chao, 2007; Jain, 2017; Punnoose & Ajit, 2016; Saradhi & Palshikar, 2011; Yigit & Shourabizadeh, 2017)
Job Satisfaction	(Jain, 2017; Punnoose & Ajit, 2016; Yigit & Shourabizadeh, 2017)
Education Level	(Chang, 2009; Dutta & Bandyopadhyay, 2020; Jain, 2017; Punnoose & Ajit, 2016; Saradhi & Palshikar, 2011; Yigit & Shourabizadeh, 2017; Zhao et al., 2019)
Marital Status	(Chang, 2009; Jain, 2017; Punnoose & Ajit, 2016; Yigit & Shourabizadeh, 2017)
Location	(Alamsyah & Salma, 2018; D & B, 2013; Dutta & Bandyopadhyay, 2020; Hong & Chao, 2007; Saradhi & Palshikar, 2011)
Stock Option Level	(Dutta & Bandyopadhyay, 2020; Jain, 2017; Yigit & Shourabizadeh, 2017)
Number of Promotions	(Dutta & Bandyopadhyay, 2020; Punnoose & Ajit, 2016; Yigit & Shourabizadeh, 2017)
Department/Team	(Alamsyah & Salma, 2018; Chang, 2009; D & B, 2013; Hong & Chao, 2007; Saradhi & Palshikar, 2011; Zhao et al., 2019)
Performance Rating	(Dutta & Bandyopadhyay, 2020; Shankar et al., 2018; Yigit & Shourabizadeh, 2017; Zhao et al., 2019)
Salary Level	(Alamsyah & Salma, 2018; Chang, 2009; D & B, 2013; Dutta & Bandyopadhyay, 2020; Hong & Chao, 2007; Jain, 2017; Punnoose & Ajit, 2016; Zhao et al., 2019)
Environmental Satisfaction	(Dutta & Bandyopadhyay, 2020; Shankar et al., 2018; Yigit & Shourabizadeh, 2017)

4.2.2 Professional Assessment

After recording the significant predictors from literature, interviews with professionals in industry were conducted to get a better perspective from churn trends that were happening within their own teams. Two managers from different industries with knowledge on data mining models and churn predication were consulted for their experience with employee churn. Along with confirming the attributes found significant in literature, each manager also selected a few other possible indicators of churn. One manager felt that an employee's decision to leave a company was heavily dependent on their micro-culture within their team. Asking questions about the number of ideas that the employee had proposed and felt were fully discusses, relative amount of empathy and support felt from their boss, freedom to choose your path within your career, and opportunities to learn something new was thought to provide better insight into the employee's likelihood of churn. Ultimately, they had found that working on a team with similar values, motivations, and professional goals helped enable retention.

The second manager had a different experience with employee churn, citing burnout and issues with management as the leading causes for an employee to leave the company. His suggestion was to assess the employee's engagement with the company and their team, their review and rating of management, normalized number of hours worked per week, and time since last promotion. After discussing these suggested measurements more thoroughly, the amount of vacation days used in the last six months was thought to also be a quantifiable way to measure employee burnout, thus helping in churn prediction.

4.2.3 Company's Assessment

The company collected data of their employees based on their own assessment of research significance and professional findings. This dataset can be split into roughly three categories: Employee Demographic, Job Position/Description, and Derived Variables (see Table 3).

Table 3: Attributes from Dataset Collected by Company

Employee Demographic	Job Position/Description	Derived Variables
1. Age 2. Gender 3. Tenure 4. Permanent Tenure 5. Manhour Training 6. Culture Results 7. Performance Score 8. Behavior Score 9. Number of Peers Promotions 10. Number of Employee's Promotions 11. Days Since Last Promotion 12. Education Level 13. University Ranking	1. Position Category 2. Division 3. Division Category 4. Employee Job Level 5. Job Level 6. Employee Type 7. Region 8. Salary Level	1. Age Normalized 2. Tenure Normalized 3. Permanent Tenure Normalized 4. Total Manhour/Total Tenure 5. Performance Score Normalized 6. Behavior Score Normalized 7. Promotion/Tenure

Many of the included attributes from the dataset have a significance in churn prediction based off the literature; the excess attributes are largely within the job position/description subset of the data. These attributes all give insight into what position the employee holds; they all really tell the same information about the employee just at varying levels of detail. It is likely that only one of the attributes within job

position/description will be used, depending on how high it is correlated with the employee's churn.

4.3 PREPROCESSING

The original dataset included 34 attributes, with 7 of these variables in some way directly dependent on other attributes. Categorical columns that contained significant information about the data objects as seen in the literature review, such as region, division category, division, and position category, were transposed into binary categories which further expanded the dimensions of the dataset. The resulting dataset had 64 columns. To reduce the working dimension of the dataset, random forest attribute selection was then used to quantify the importance of each variable in classifying the employee's churn, shown in Figures 10-15.

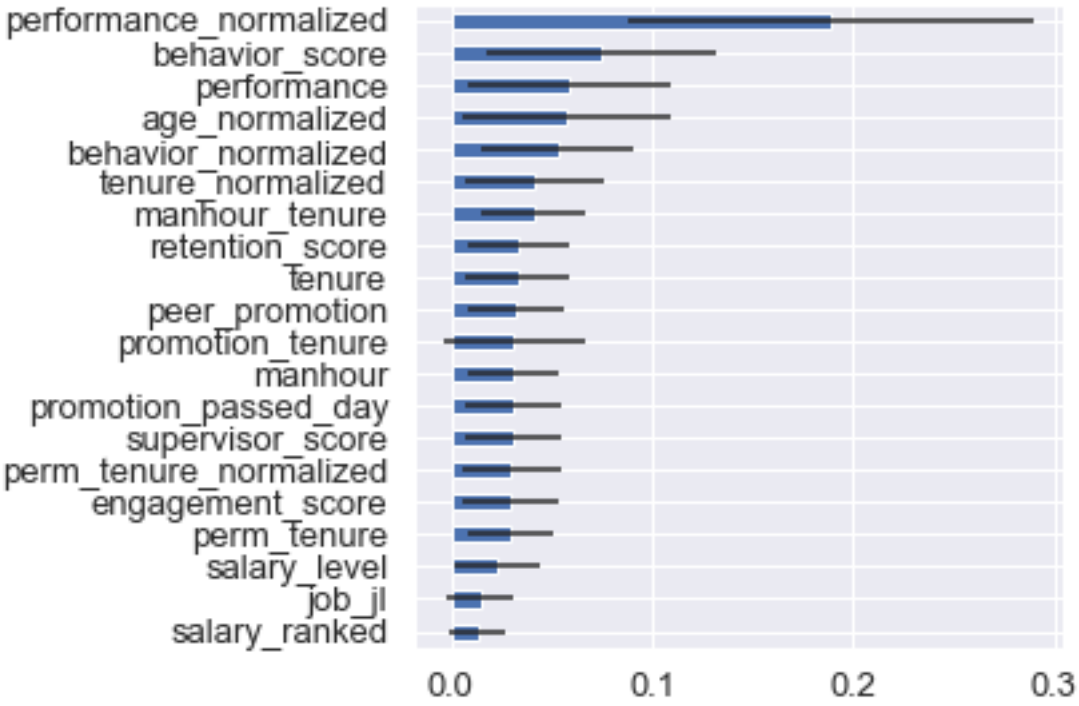

Figure 10: Random Forest Attribute Selection without Age

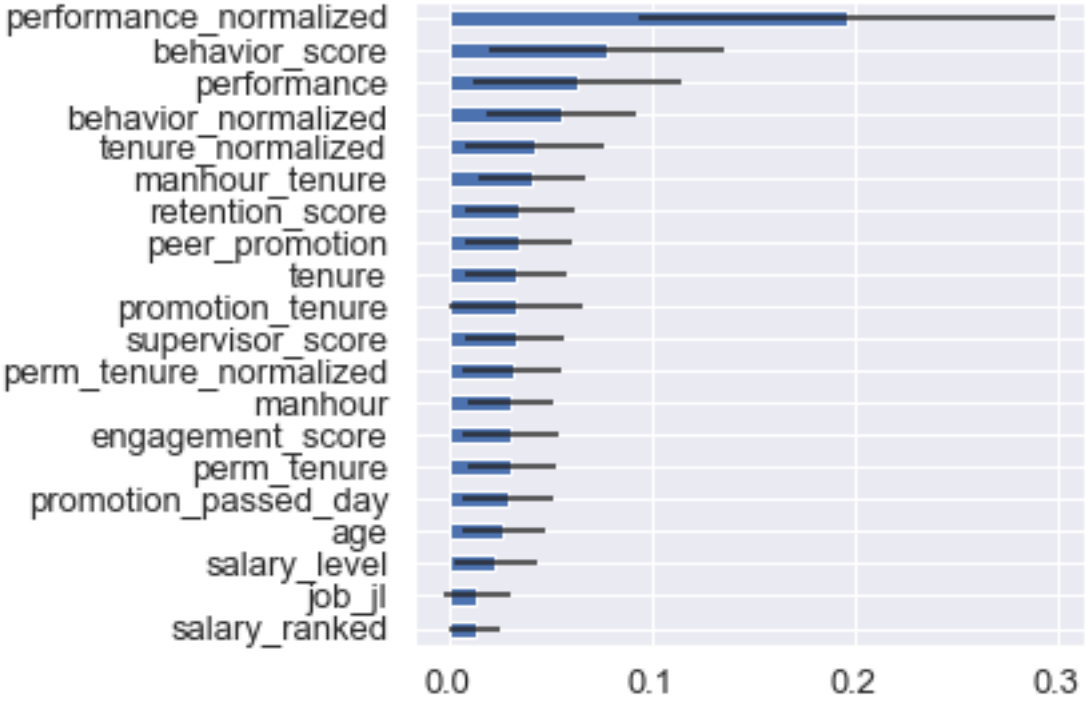

Figure 11: Random Forest Attribute Selection without Age Normalized

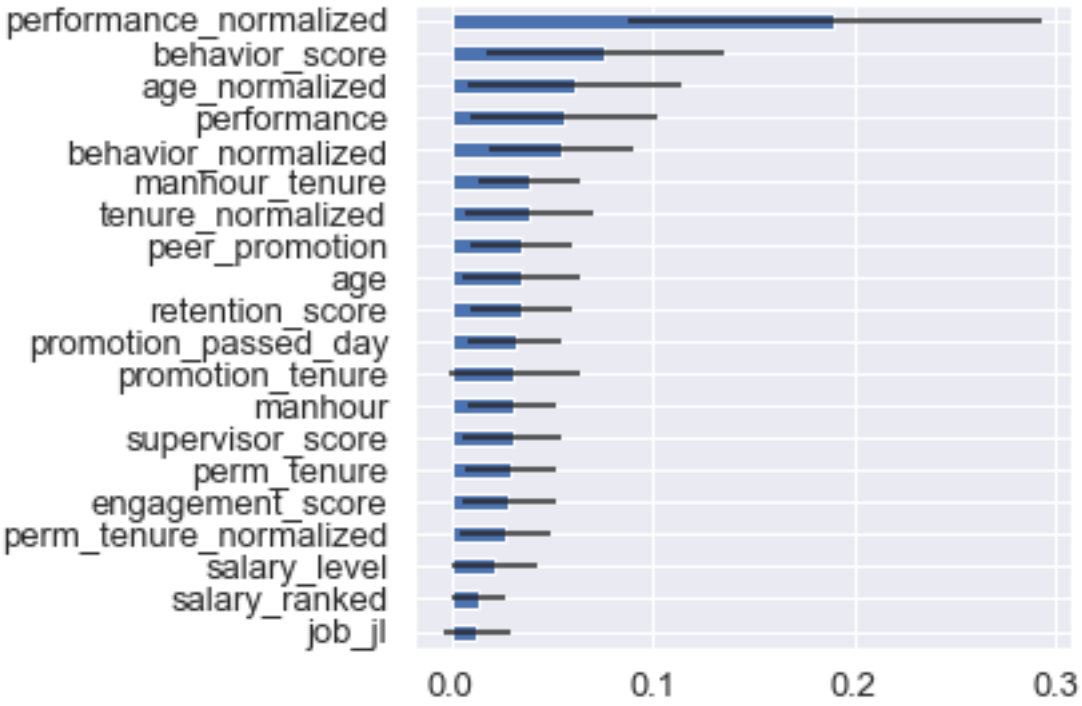

Figure 12: Random Forest Attribute Selection without Tenure

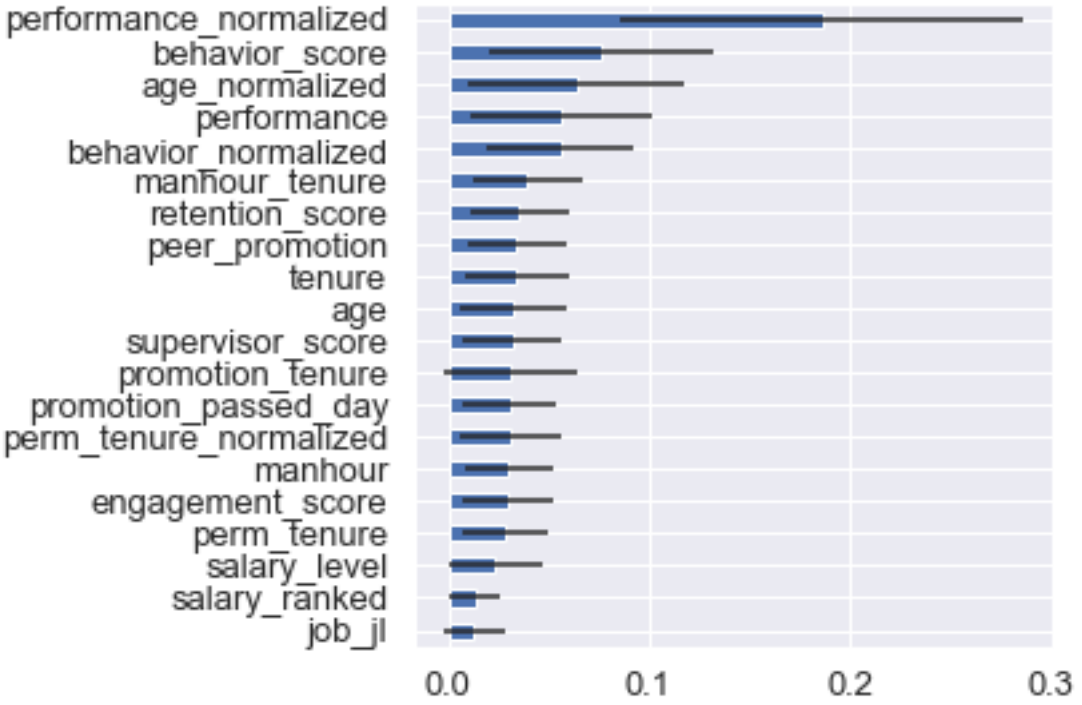

Figure 13: Random Forest Attribute Selection without Tenure Normalized

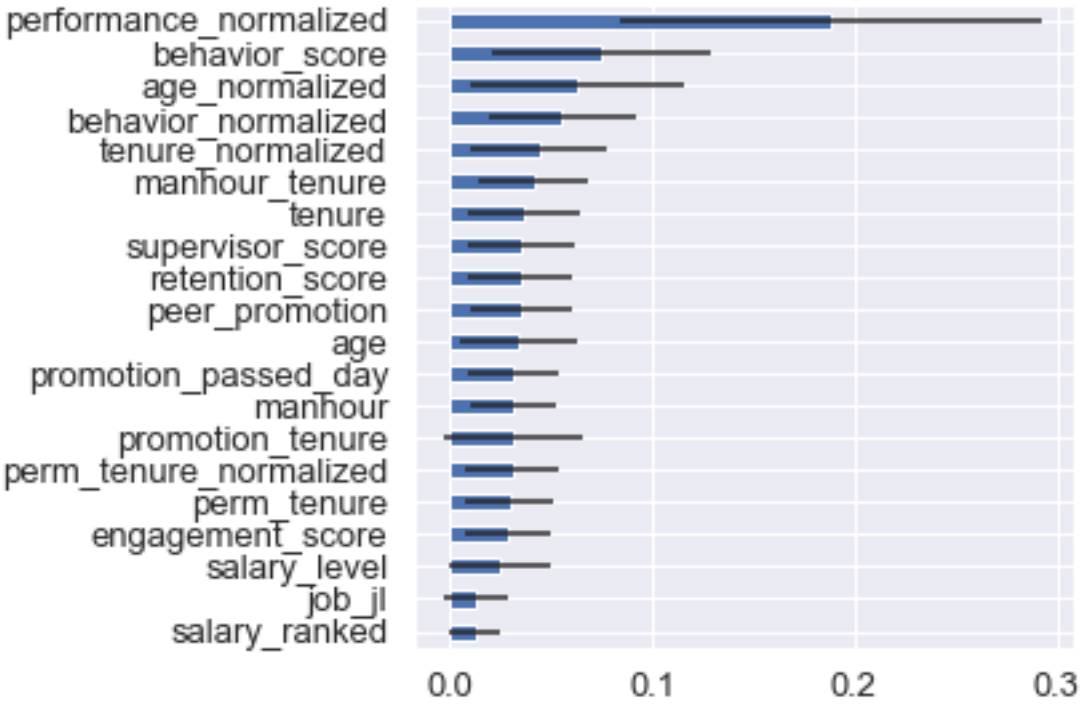

Figure 14: Random Forest Attribute Selection without Performance Score

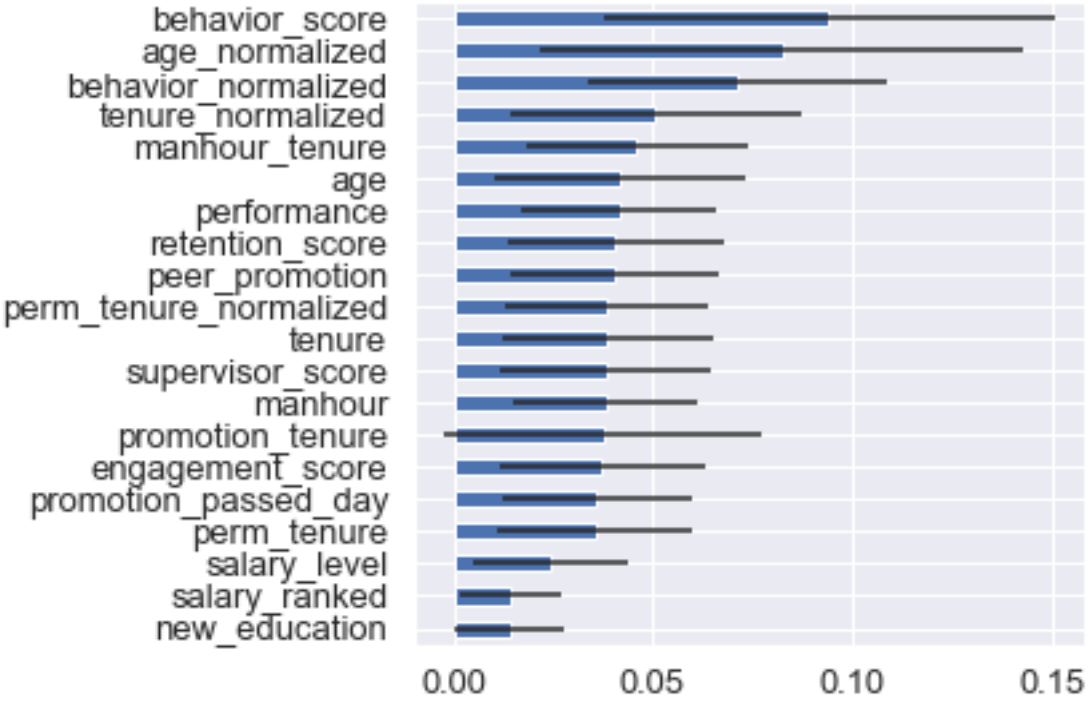

Figure 15: Random Forest Attribute Selection without Performance Score Normalized

Random forest (RF) creates iterations of decision tree classifier and records the effectiveness of each attribute in improving the split criterion in the corresponding splits in the tree. The higher the calculated importance of an attribute, the more successful it is in correctly classifying churn. Figures 1-6 show the twenty most important attributes in churn prediction for the dataset given a highly correlated variable is missing from the analysis. RF was repeated six times to determine which variable should be retained for highly correlated pairs age vs. age normalized, tenure vs. tenure normalized, and performance vs. performance normalized. In each analysis, age normalized, tenure normalized, and performance normalized were more important in classification than their non-normalized pair.

To further analyze the relationship between variables, the correlation values were calculated and values greater than 0.7 were evaluated further to reduce data redundancy. A correlation heatmap was constructed, depicted in Figure 16.

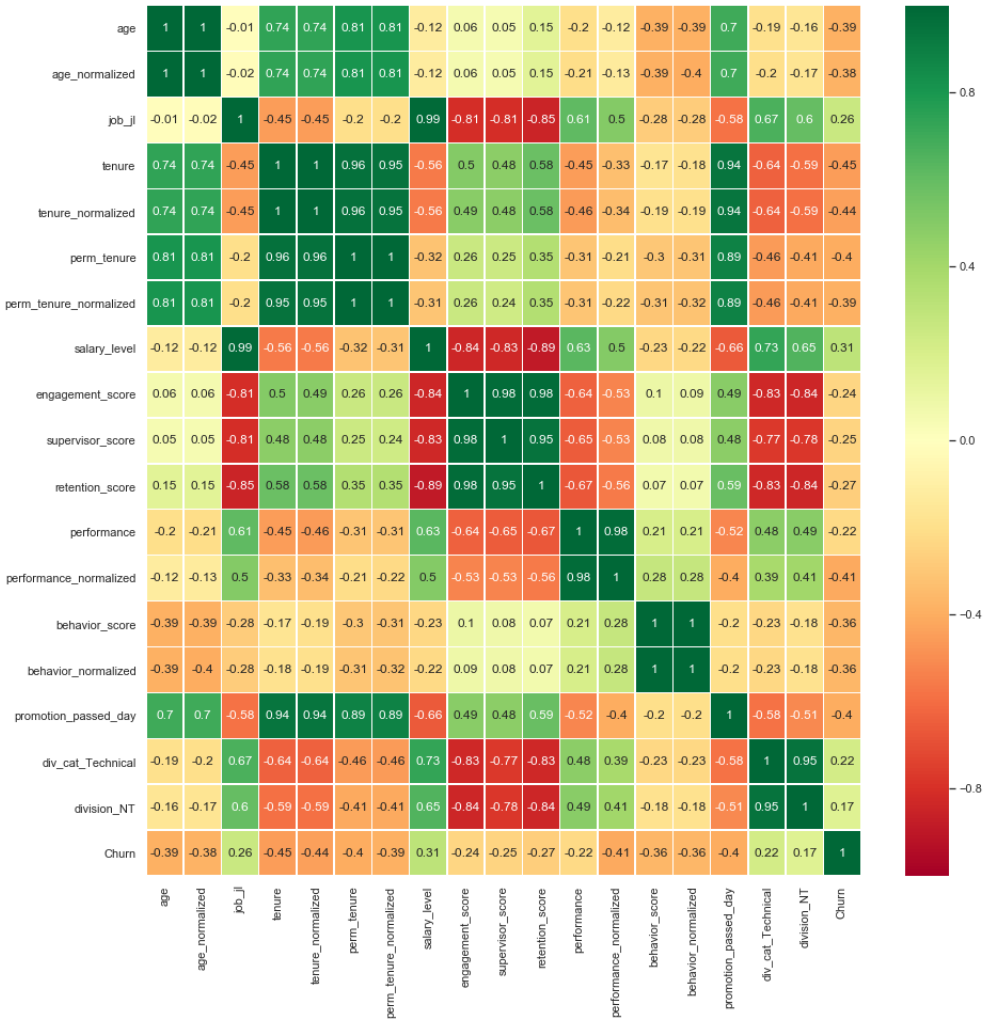

Figure 16: Correlation Heatmap for Strongly Associated Variables

Only one attribute in a highly correlated cluster should be retained in the usable dataset to avoid redundancy and overfitting in the model. Table 4 details which attributes had high correlation with others, and which were retained for classification. The decision of which attributes to retain was guided by the results shown in Figures 10-15 as well as the correlation values in Table 4.

Table 4: Highly Correlated Cluster Summary

Retained Attribute	Highly Correlated Attributes	Correlation
Age normalized	Age	1.00
	Permanent tenure	0.81
	Permanent tenure normalized	0.81
Tenure normalized	Age	0.74
	Tenure	1.00
	Permanent tenure	0.96
	Permanent tenure normalized	0.95
Salary level	Job level	0.99
Performance normalized	Performance	0.98
Behavior score	Behavior score normalized	1.00
Division category – Technical	Division – NT	0.95

Values were only dropped from the usable dataset if they were in some way obviously related to other variables in the dataset and had a high correlation value. For example, tenure and tenure normalized are directly related, but salary level and culture scores do not immediately imply a direct relationship. The attributes where the relationship was indirect were retained for further analysis.

Recursive feature elimination (RFE) was applied to the dataset to gain a better understanding of how the variables interact with one another in churn classification. The goal of RFE is to iterate over the possible combinations of the given attributes and continue to consider smaller attribute sets until only the most important attributes remain. Using 10-fold cross validation and decision tree classifier, RFE is tested with required attributes ranging from 2-59. The output of each iteration is a Boolean mask equal to the input dimensions of attributes; this portrays which attributes are retained for classification in this trial. Each trial adds the next best attribute for classification in order to increase recall, until the final trial where all attributes are used. The recall score for each model

was recorded and box plots of the distributions for every level of attribute inclusion was created, as can be seen in Figure 17.

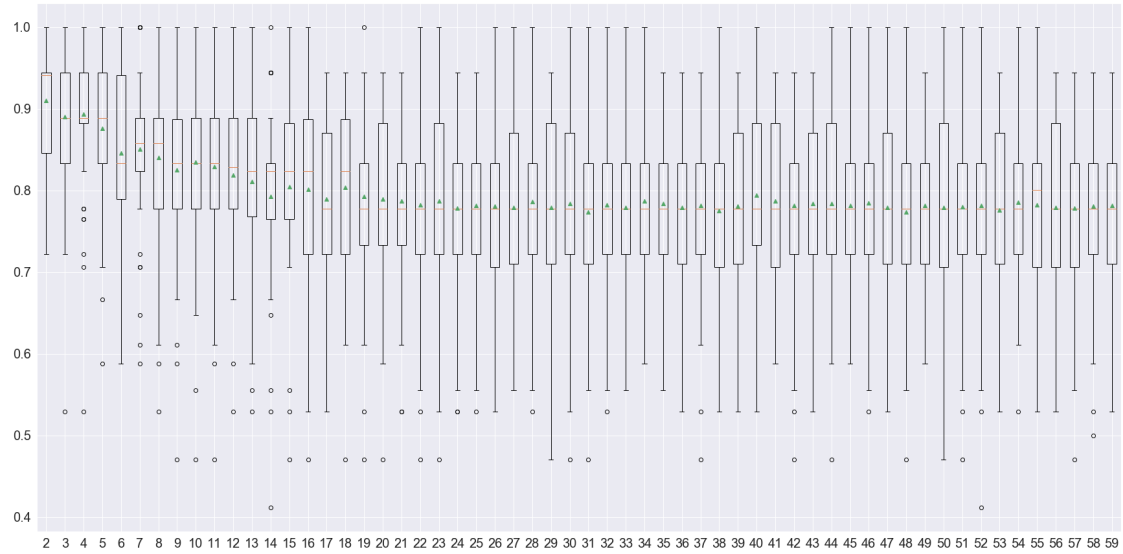

Figure 17: Boxplots of Recall Score for Incremental Inclusion of Attributes

The intervals show there is no significant difference in recall score from the inclusion of more attributes. However, there is a trend of higher average recall scores in the initial, smaller trials, implying that the inclusion of these initial variables has a greater impact on churn performance than the subsequent attributes.

4.3.1 Missing Values

From the dataset, culture scores, performance score normalized, and behavior score had missing values. Figure 18 shows a positive correlation between missing values and the churning class.

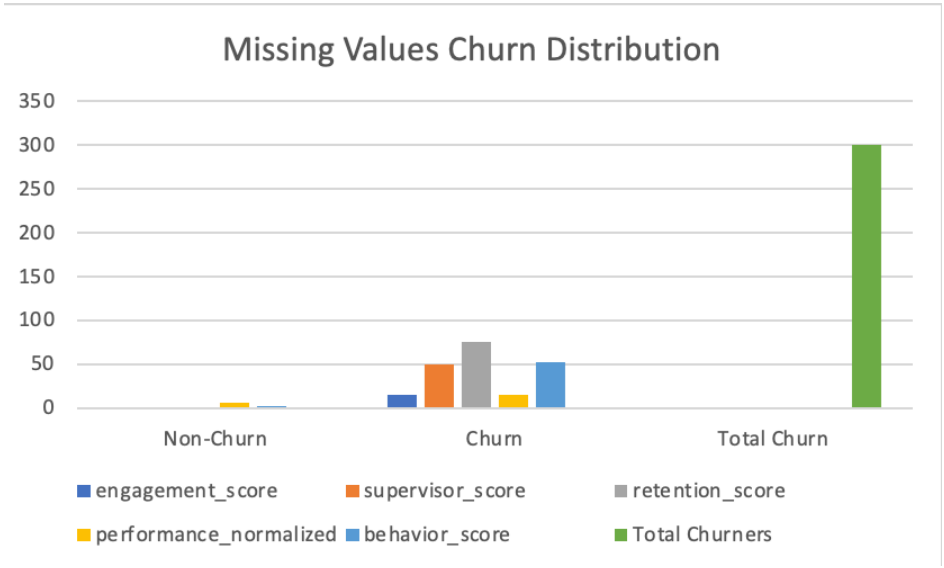

Figure 18: Distribution of Churn by Missing Values

The total number of churners in the full dataset was included in Figure 18 to better understand the correlation between missing values and churn. Further analysis showed that these missing values correlated to employees just starting their career at the company; the data objects tended to have lower tenure than the distribution from the entire dataset. A representative from the company confirmed that the data attributes culture scores, performance scores, and behavior scores are calculated semiannually; therefore, if an employee has worked for the company for less than six months then there is a higher likelihood the data object will have at least one of these values missing. Figure 19 shows the relationship described above; the null data had a lower skewed tenure distribution than when compared to the full dataset.

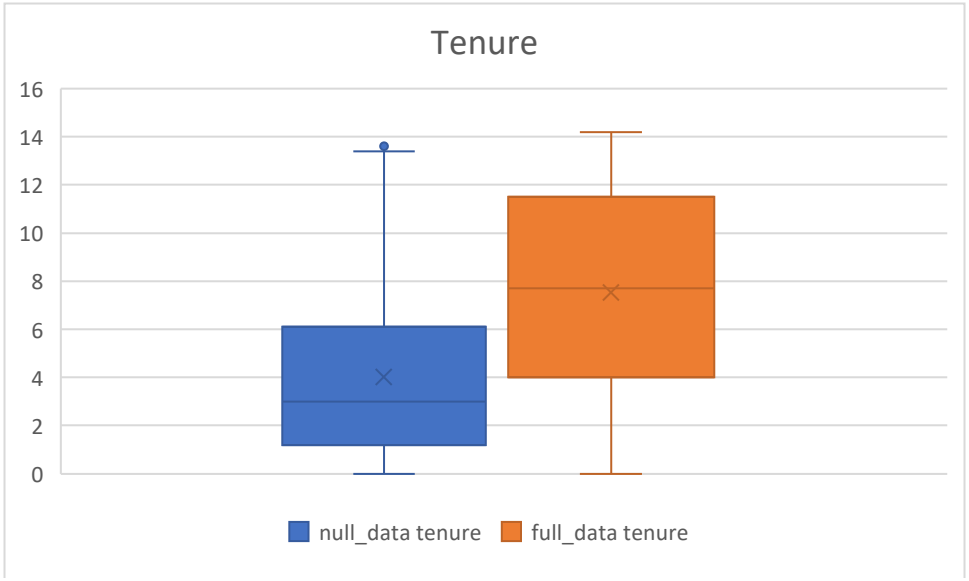

Figure 19: Boxplot of Tenure Distribution from Null Data Compared to Full Dataset

Within the null data, 67 data objects had two or more missing values; these were dropped from the usable dataset to avoid over generalizing the employee's information. There are two options to handle these missing values: impute the missing values using prediction or summary statistics or drop data objects with more than one missing value. Imputing the missing values will incorporate more bias into the data objects, which could reduce overall classification of churning employees. However, if there is a strong association between newer employees and their likelihood to churn, removing these data objects could also negatively impact churn classification for low-tenured employees. Revisiting the goals for this research, the company has determined that the higher priority should be on classifying churn in employees that have been with the company for longer as they have already demonstrated their value and the company has fully invested in the employee. Dropping the data objects with multiple null values aligns more with the company's objective, and as can be seen from Figure 20, the remaining null data objects have a similar tenure distribution as the entire dataset. This further validates the original

assumption that newer employees will have multiple missing values due to not being with the company when different scores are calculated.

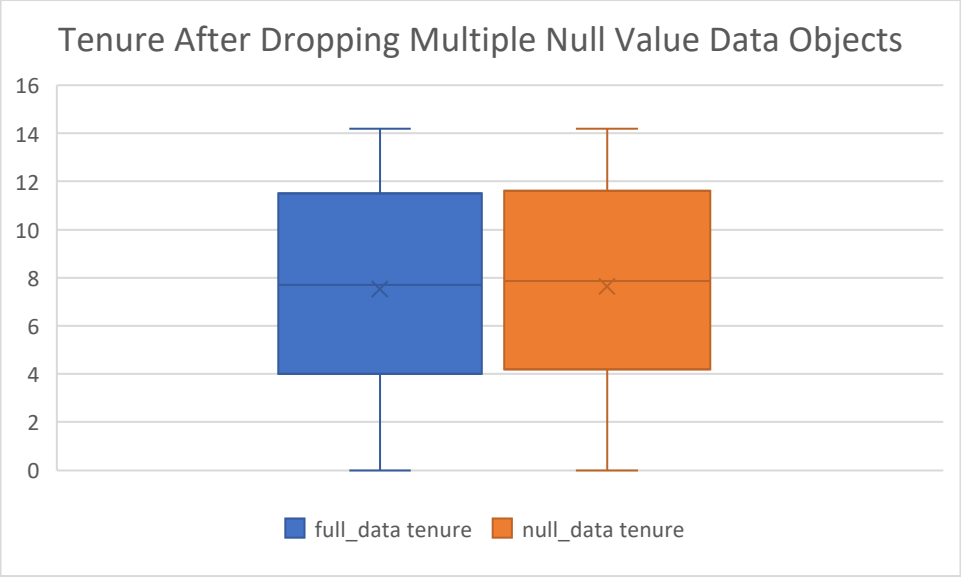

Figure 20: Tenure After Dropping Multiple Null Value Data Objects

After dropping the data objects with more than one null value, the frequency of missing values greatly reduced to 24 missing values for culture retention score, 4 for performance normalized, and 36 for behavior score. Looking back at the full dataset, all three culture score attributes are highly correlated with each other (seen in Table 5).

Table 5: Culture Scores Correlation Values

	Engagement	Supervisor	Retention	Churn
Engagement	1.00	0.98	0.98	-0.24
Supervisor	0.98	1.00	0.95	-0.25
Retention	0.87	0.95	1.00	-0.27
Churn	-0.24	-0.25	-0.27	1.00

For the classification models, only one culture score will likely be retained as they all account for a majority of the same variance in an employee's churn. Referring back to Figure 10-15, all three culture scores have approximately the same importance score with

the variance in importance for culture supervisor less than that of the other attributes. Rather than predicting values for culture retention and incorporating bias into the dataset, this attribute will be dropped as a majority of the variance is accounted for in the other culture variables with similar importance in classifying churn.

Regression prediction was tested for imputing behavior and performance normalized, but both tests resulted in a low coefficient of determination (less than 0.10), so regression imputation was not a viable option. Since both of these variables are significant in churn prediction, the median value for the data objects churn class is imputed for each measure.

4.3.2 Outlier Analysis

There are two methods used to address outliers: univariate and multivariate. Univariate analysis identifies outliers by addressing each column individually, while multivariate analysis takes all column values into account and identifies entire data objects that are significant outliers from the population. This research includes an initial univariate assessment on the individual attributes as well as a multivariate analysis on the data objects as a whole to pinpoint which data objects are likely to negatively impact the classification algorithms.

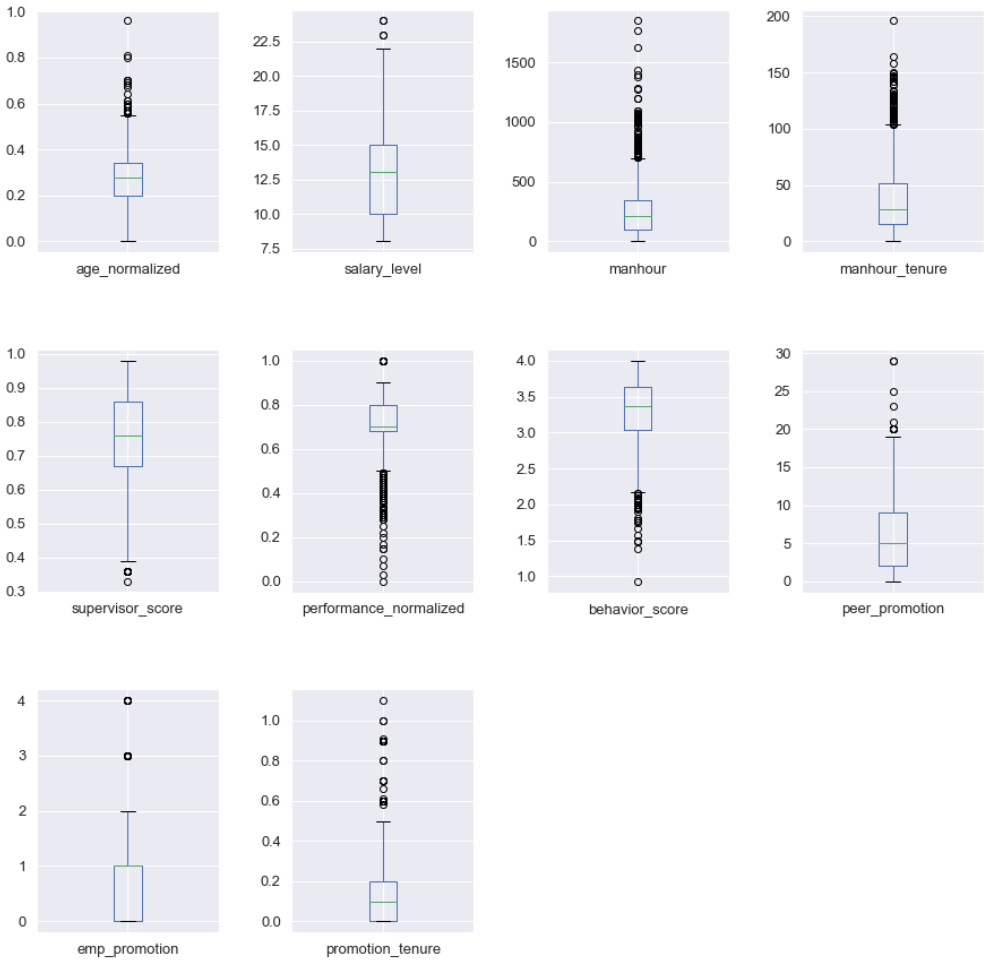

Figure 21: Boxplot of Attributes with Extreme Values from Full Dataset

As shown in Figure 21, age normalized, salary level, manhour, manhour/tenure, culture supervisor, performance normalized, behavior score, number of peer promotions, number of employee promotions, and promotion/tenure all had a number of extreme values. The indices for the data objects with extreme values were saved to include in the multivariate outlier analysis. The Mahalanobis distances for the data objects with extreme values were calculated and compared to determine which data objects were negatively affecting the dataset. Mahalanobis distance compares the individual data object to the distribution of the dataset as a whole. A threshold distance value of 73.311 was determined based on the

chi-squared distribution and the number of parameters in the usable dataset. The distance measures were calculated for the extreme data objects, and 24 data objects had distances greater than the threshold value; of the 24 outliers, only 5 belonged to the churning class. These were dropped from the usable dataset to avoid skewing the algorithms and incorporating bias into the dataset. The remaining extreme values were unaltered since their distance from the full dataset was not seen as significant enough to skew the classification results.

4.4 SUMMARIES AND STATISTICS

The attributes retained for churn classification and the data type, important parameters, and descriptions can be found in Table 6. The resulting dataset is comprised of 2072 rows and twenty columns: nineteen independent and one dependent attribute. 230 data objects are classified as churners, resulting in a 11.10% churn rate.

Table 6: Churn Data Summary Statistics

Attribute Name	Description	Data Type	Stats
Performance Normalized	Direct supervisor's evaluation of the employee's quarterly performance, normalized by the data from that quarter and year	Continuous Float	Mean = 0.712 Median = 0.7 St. Dev. = 0.122
Behavior Score	A score given by higher ups to portray an employee's behavioral competencies	Continuous Float	Mean = 3.31 Median = 3.36 St. Dev. = 0.42
Manhour/ Tenure	Ratio of total number of training hours to employee tenure	Continuous Float	Mean = 36.30 Median = 28.805 St. Dev. = 27.66

Peer Promotions	Number of peers of the employee that have been promoted	Integer	Mean = 5.96 Median = 5.0 St. Dev. = 4.10
Culture Score - Supervisor	Culture score for the employee's general satisfaction of their direct supervisor	Continuous Float	Scale = 0-1 Mean = 0.75 St. Dev. = 0.12
Tenure	Amount of time in years the employee has worked for the company at the time the data was collected	Continuous Float	Mean = 7.65 Median = 7.9 St. Dev. = 4.03
Promotion/ Tenure	Ratio of the number of promotions an employee has received and their tenure with the company	Continuous Float	Mean = 0.13 Median = 0.10 St. Dev. = 0.14
Days Since Last Promotion	Number of days passed since the last promotion the employee has received, if the employee has not been promoted value will be equal to their tenure	Integer	Min = 1 Max = 5135 Mean = 1744.69 Median = 1551.00 St. Dev. = 1207.09
Manhour	Total number of training hours an employee has had since starting for the company	Continuous Float	Mean = 246.46 Median = 209.00 St. Dev. = 198.64
Age	Age of the employee at the time the data was collected, for churn employees age will equal the age they left the company	Integer	Mean = 35.61 Median = 36.00 St. Dev. = 4.58
Salary Level	Rating of compensation and salary level	Integer Rank	Range = 8-24 Median = 13.00
Culture Score - Engagement	Culture score for the employee's level of engagement with the company and their environment	Continuous Float	Scale = 0-1 Mean = 0.74 St. Dev. = 0.11

Salary Level Ranked	Salary level ranked within the employee's job level	Integer Rank	Range = 1-5 Median = 3.00
Education Level	Highest level of education employee has received: HS Diploma, Associate, Bachelor, Master, PhD	Integer Rank	(1) Diploma = 58 (2) Associate = 227 (3) Bachelor = 1253 (4) Masters = 529 (5) PhD = 5
University Ranking	Indicator of whether or not the employee went to a top ranked Iranian university	Binary	# Top Uni = 353 # Not Top Uni = 1719
Employee Promotions	Number of promotions the employee has received while employed for the company	Integer	Min = 0 Max = 4 Median = 1
Gender	Gender of the employee	Binary	Male = 1313 Female = 759
Finance Division	Indicator of whether or not the employee's position falls within finance	Binary	# in Finance = 149 # Not in Finance = 1923
Admin Position Category	Indicator of whether or not the employee is in an administrative position	Binary	# Admin = 753 # Not Admin = 1319
Churn	Indicator of employee's churn	Binary	# Churners = 230 # Non-Churners = 1842

5. DESIGN OF EXPERIMENT

5.1 K-FOLD CROSS VALIDATION

The objective of this case study is to identify the optimal combination of undersampling and data mining algorithm necessary to classify employee churn. In order to draw significant conclusions, k-fold cross validation will be used to control extraneous variation. This study will include ten trials, each fold being a different one-tenth of the original dataset with 90% belonging to the train dataset and 10% reserved as a test dataset (see Table 7). Within each trial, all treatments will have one replicant and results will be averaged across treatments.

Table 7: 10-Fold Cross Validation

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10
Test Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set
Train Set	**Test Set**	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set
Train Set	Train Set	**Test Set**	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set
Train Set	Train Set	Train Set	**Test Set**	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set
Train Set	Train Set	Train Set	Train Set	**Test Set**	Train Set	Train Set	Train Set	Train Set	Train Set
Train Set	Train Set	Train Set	Train Set	Train Set	**Test Set**	Train Set	Train Set	Train Set	Train Set
Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	**Test Set**	Train Set	Train Set	Train Set
Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	**Test Set**	Train Set	Train Set
Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	**Test Set**	Train Set
Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	Train Set	**Test Set**

5.2 CLUSTERING AND UNDERSAMPLING

The usable dataset has a total of 2072 non-churners and 230 churners; only about 11.10% of the data belongs to the class of interest. For each trial, the full dataset is split into

training and testing subsets depending on the fold being used as the testing set, as shown in Figure 22.

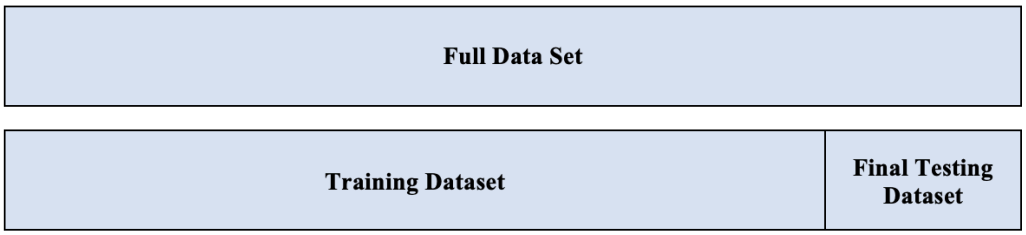

Figure 22: Training and Test Dataset Split for a Single Trial

The resulting training dataset has approximately the same proportion of churners as the full dataset. A clustering algorithm called K-means is then used as a preprocessing step before under sampling. K-Means partitions the dataset into 'k' specified clusters based on their sameness. In order to balance the dataset, there has to be an equal number of non-churning clusters as there are churners in the training dataset; for each trial, k was set to equal the number of churners and the data was partitioned into as many subgroups. The clusters average value for each attribute was calculated and recorded as a new data object to be used later for under sampling. Table 8 shows a summary of the number of clusters as well as the smallest and largest cluster by trial.

Table 8: k-Means Summary by Trial

Trial	k	Smallest Cluster Size	Largest Cluster Size
1	207	1	24
2	208	1	32
3	208	1	35
4	208	1	38
5	208	1	31
6	208	1	47
7	208	1	28
8	208	1	33
9	208	1	30
10	208	1	27

kNN is a classification method that relies on the "nearest neighbors", or most similar data object, of a data point to classify the new object. The algorithm does not require any training and does not perform as well as classification algorithms mentioned earlier, such as decision tree or MLP. "k" determines the number of neighbors to a data object that are looked to for classification; the data object's classification is determined by the majority class from its k most similar data objects. For each cluster, the new data object that was saved as the average of all attributes within the cluster was compared within the corresponding cluster to find the actual data object that it was most like. Rather than classifying the aggregate data object, the k most similar data objects were used as representatives of the cluster. The value of k will greatly affect the balance in the dataset, and therefore the performance and effectiveness of the classification algorithm. To find the optimal values of k, the training data set was split into training and testing subsets to judge the accuracy of the different levels of k in classifying data objects into the correct cluster. Accuracy is used to assess the best number of data objects to represent the corresponding cluster for predicting employee churn because it is assumed that the more

similar the data objects are to its cluster, the better a representation of the entire dataset. A better summary of the population data would then lead to a better churn classification performance. k of one through fifteen was tested for each block and the accuracy at classification was compared. In general, as k increases, accuracy decreases. k of one, three, and five have the highest average accuracy across all trials and are selected for future comparison. Three datasets are made, each using a different value of k to sample the nearest neighbors of each cluster. A fourth dataset using random sampling to balance the classifiers was also created to serve as a baseline for the experiment, and a fifth dataset that didn't use any undersampling method and instead contained the entire train dataset from the trial was used as a control (see Figure 23). This step is repeated per trial, resulting in forty undersampled datasets and ten control datasets.

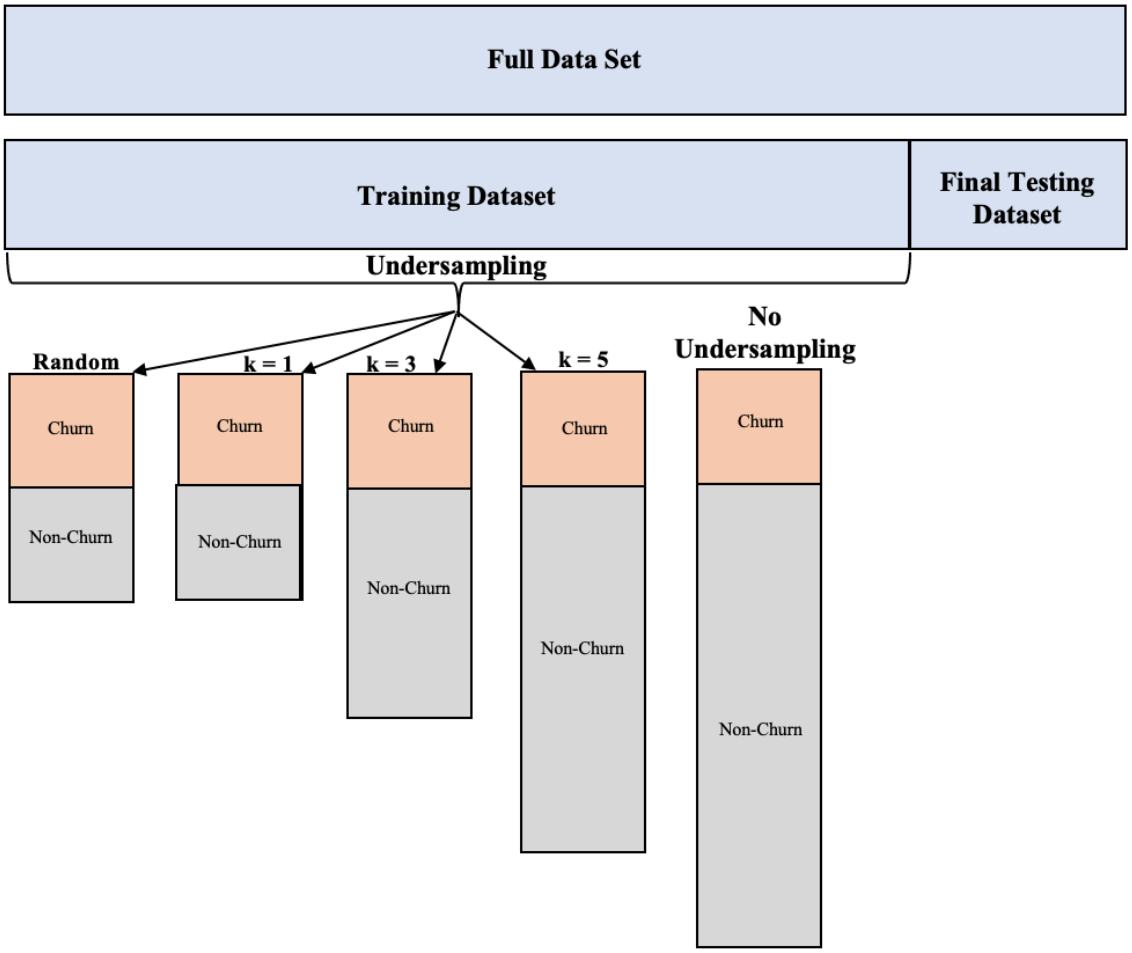

Figure 23: Representation of a Trial's Undersampling from the Training Dataset

In applicable datasets, after all nearest neighbors are chosen for each cluster, the data set is filtered to only allow an employee number to appear once in the dataset, even if the same data object was chosen as a nearest neighbor for multiple clusters. These datasets will be evaluated and compared to a randomly sampled balanced subset and a non-balanced subset to determine the best under sampling method. Table 9 is a summary of each trial's resulting balance by undersampling method.

Table 9: Summary of Final Model Datasets

Trial	Undersampling Method	Total Non-Churners	Unique Non-Churners	Total Churners	Churn Rate
1	kNN, k = 1	207	207	207	50.00%
	kNN, k = 3	621	596	207	25.78%
	kNN, k = 5	1035	900	207	18.70%
	Random	207	207	207	50.00%
	None	1655	1655	207	11.12%
2	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	595	208	25.90%
	kNN, k = 5	1040	905	208	18.69%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
3	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	601	208	25.71%
	kNN, k = 5	1040	911	208	18.59%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
4	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	606	208	25.55%
	kNN, k = 5	1040	906	208	18.67%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
5	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	612	208	25.37%
	kNN, k = 5	1040	915	208	18.52%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
6	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	600	208	25.74%
	kNN, k = 5	1040	903	208	18.72%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
7	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	604	208	25.62%
	kNN, k = 5	1040	920	208	18.44%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
8	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	611	208	25.40%
	kNN, k = 5	1040	922	208	18.41%
	Random	208	208	208	50.00%

	None	1655	1655	208	11.16%
9	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	604	208	25.62%
	kNN, k = 5	1040	912	208	18.57%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%
10	kNN, k = 1	208	208	208	50.00%
	kNN, k = 3	624	611	208	25.40%
	kNN, k = 5	1040	929	208	18.29%
	Random	208	208	208	50.00%
	None	1655	1655	208	11.16%

5.3 ALGORITHM SELECTIONS

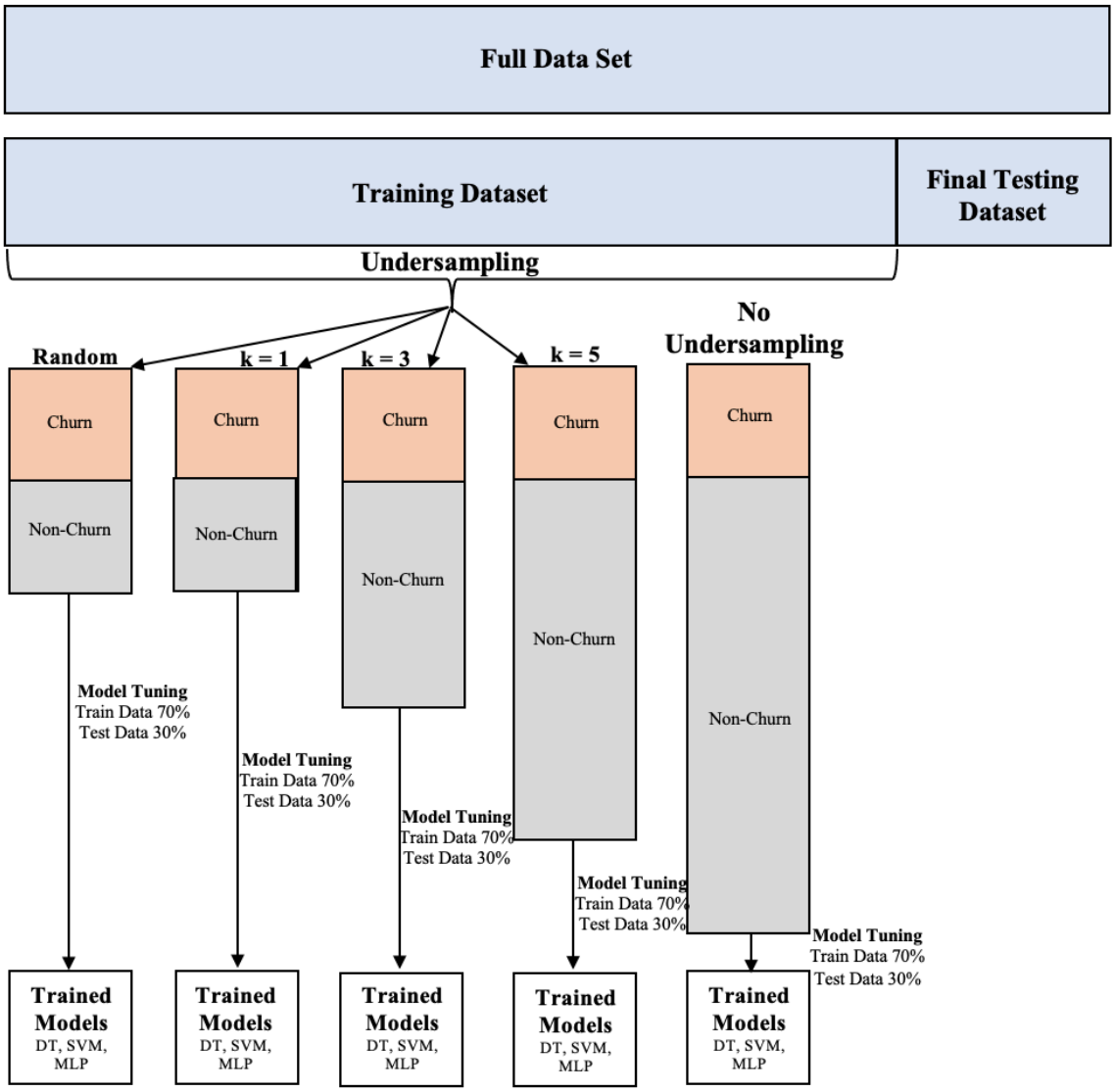

Figure 24: Visualization of Trial's Tuning Models from Sampled Datasets

Using the fifty training datasets, three algorithms will be applied and tuned in order to find the superior classifier of employee churn (see Figure 24). The algorithms selected for this study are decision tree, support vector machine, and multi-layer perceptron. While all very different, these three algorithms were chosen as they each bring a different feature to classification that the others lack. Decision Tree's output is highly interpretable, and easy

to pull decision rules from and apply changes to business practices in order to decrease employee churn. Support Vector Machines have proven to be significant classifiers and successful churn predictors in literature. Multi-Layer Perceptron excels with identifying non-linear relationships between dependent and independent variables, increasing classification performance with more complex enterprise problems. The resulting tuned algorithms from all datasets will be tested using the original validation data from the entire dataset.

5.4 ALGORITHM TUNING

5.4.1 Decision Tree

Decision tree tuning involves finding a maximum depth (MD), minimum impurity decrease (MID), and minimum sample split (MSS) for each sample set. Without specifying these metrics, the algorithm would be overfit to the training dataset and would not be successful in classifying churn from new data. To find the best values, arbitrary ranges for each input were tested using random grid search. Based on this initial output, the values were further redefined with new ranges and from there optimal values were chosen. The optimal values were chosen based on their effect on the recall score for the model. Recall was the chosen metric as it prioritizes the correct classification of churners, which is of more importance to the company. This step was repeated for each dataset and the results recorded and applied during model construction (see Table 10).

Table 10: Decision Tree Tuning Values

	Undersampling														
	k = 1			k = 3			k = 5			Random			None		
Trial	**MD**	**MID**	**MSS**	**MD**	**MID**	**MSS**	**MD**	**MID**	**MSS**	**MD**	**MID**	**MSS**	**MD**	**MID**	**MSS**
1	1	0.005	5	7	0.0002	16	1	0.0025	45	16	0.005	5	18	0.0001	6
2	4	0.0005	17	48	0.0005	5	1	0.0025	45	5	0.005	21	4	0.0001	25
3	9	0.005	5	25	0.001	6	8	0.0001	40	4	0.0001	25	25	0.0001	38
4	3	0.0001	35	9	0.0001	25	1	0.005	5	7	0.0001	17	16	0.0001	35
5	4	0.01	5	1	0.005	5	5	0.0001	54	8	0.005	8	7	0.0002	40
6	22	0.0005	11	1	0.0025	5	3	0.005	39	5	0.01	21	7	0.0001	44
7	4	0.0075	6	45	0.0006	5	1	0.0002	5	4	0.0125	21	9	0.0001	55
8	5	0.01	5	18	0.005	7	7	0.0025	31	19	0.0062	6	10	0.0001	42
9	6	0.0001	11	26	0.0002	6	29	0.0005	6	4	0.0038	5	10	0.0001	5
10	19	0.0075	5	7	0.0003	27	6	0.0001	56	4	0.005	16	6	0.0001	37

5.4.2 Support Vector Machines

There are three parameters that need to be tuned for SVM models: kernels, gamma, and weight. The kernel type specified determines the method for class detection in pattern analysis. Gamma determines the influence each data object has on the training of the hyperplane; weight defines the different weight of the data objects in classification. In order to find the best parameters, every combination of these metrics was tested and it's resulting recall on the tuning dataset was recorded. The optimal combination of tuning metrics was selected again based on the improvement of the recall score and was applied to the validation data set (see Table 11).

Table 11: SVM Tuning Values

	Undersampling														
	k = 1			k = 3			k = 5			Random			None		
Trial	**KT**	**GM**	**Wgt**	**KT**	**GM**	**Wgt**	**KT**	**GM**	**Wgt**	**KT**	**GM**	**Wgt**	**KT**	**GM**	**Wgt**
1	linear	None	auto	linear	None	scale	linear	None	auto	linear	bal.	auto	linear	None	auto
2	linear	bal.	scale	linear	None	auto	linear	None	scale	linear	None	auto	rbf	None	auto
3	linear	None	auto	linear	None	auto	linear	None	auto	linear	None	auto	linear	None	scale
4	linear	bal.	auto	linear	None	scale	linear	None	scale	linear	bal.	scale	linear	None	auto
5	linear	None	scale	linear	None	scale	linear	None	auto	linear	None	auto	linear	None	auto
6	linear	bal.	scale	linear	bal.	scale	linear	None	auto	linear	bal.	auto	linear	None	scale
7	linear	bal.	scale	linear	None	auto	linear	bal.	scale	linear	None	auto	linear	None	auto
8	linear	None	auto	linear	None	auto	linear	None	scale	linear	None	auto	linear	None	auto
9	linear	bal.	scale	linear	bal.	auto	linear	None	auto	linear	bal.	auto	linear	None	scale
10	linear	bal.	auto	linear	None	auto	linear	None	scale	linear	bal.	auto	linear	None	scale

5.4.3 Artificial Neural Networks

Artificial Neural Networks are random based algorithms; multi-layer perceptron specifically assigns random weights and interconnections between independent and dependent variables and learns from the random assignment to assign optimal weights for performance. In order to tune the random based algorithm MLP, 115 single and multi-layer network structures were tested with different activation functions and solver options. Activation functions determine how the inputs are processed at the nodes within the neural network, while the solver parameter determines the optimization method for the weights of the interconnectors. The recall of the tuning metric combinations was recorded, and the values with the highest recall were retained. Optimal hidden structure size, activation, and solver were chosen for each dataset and MLP algorithms were constructed based on these findings (see Table 12).

Table 12: MLP Tuning Values

	Undersampling														
	k = 1			k = 3			k = 5			Random			None		
Trial	**HS**	**Act.**	**SV**	**HS**	**Act.**	**SV**	**HS**	**Act.**	**SV**	**HS**	**Act.**	**SV**	**HS**	**Act.**	**SV**
1	[1]	log	lbfgs	[7]	iden.	sgd	[4]	iden.	sgd	[1,10]	log	adam	[2,9]	relu	lbfgs
2	[3,3]	log	adam	[1]	relu	adam	[4,1]	relu	lbfgs	[1,2]	log	lbfgs	[7,1]	relu	adam
3	[2]	log	lbfgs	[14]	iden.	sgd	[13]	iden.	sgd	[13]	iden.	sgd	[5,2]	relu	adam
4	[1]	log	lbfgs	[1,1]	tanh	adam	[8,1]	relu	adam	[1,1]	log	lbfgs	[6,1]	relu	adam
5	[1,1]	log	sgd	[12]	iden.	sgd	[5,5]	relu	lbfgs	[14]	iden.	sgd	[3,3]	relu	adam
6	[2]	log	lbfgs	[11]	iden.	sgd	[10]	iden.	sgd	[1,2]	log	lbfgs	[3,10]	relu	lbfgs
7	[1]	log	lbfgs	[10]	iden.	sgd	[4]	iden.	sgd	[2,9]	log	sgd	[3,3]	relu	lbfgs
8	[1,1]	log	lbfgs	[3]	iden.	sgd	[15]	iden.	sgd	[1,8]	log	adam	[6,2]	relu	lbfgs
9	[4]	log	adam	[1,1]	relu	adam	[7]	iden.	sgd	[1]	log	lbfgs	[4,4]	relu	lbfgs
10	[2,6]	log	adam	[8]	iden.	sgd	[10,1]	relu	adam	[2,10]	log	lbfgs	[7,2]	relu	lbfgs

6. RESULTS

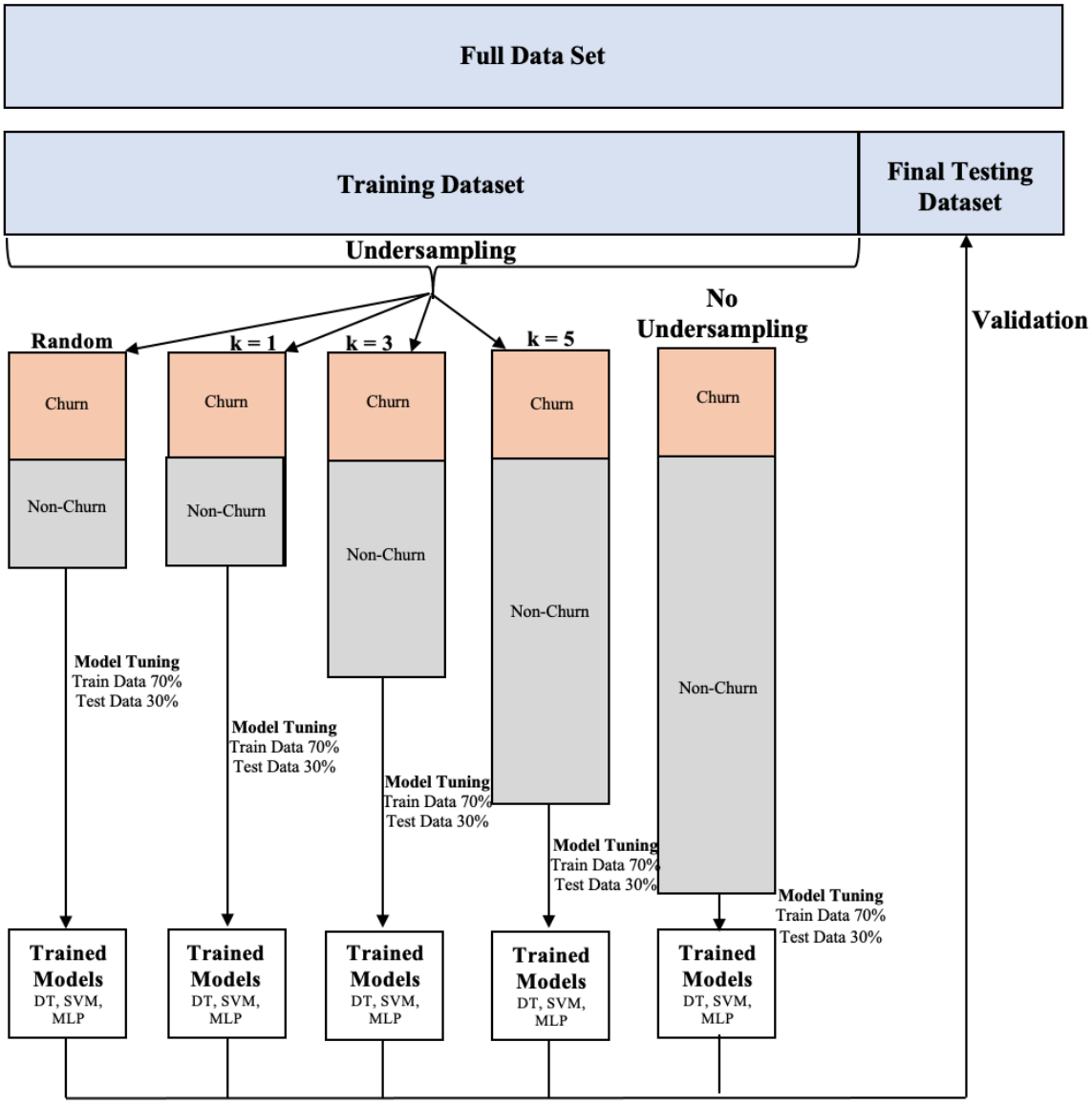

Figure 25: Visualization of Full Data Model Flow

After defining the algorithm metrics for each dataset, all models were then validated using the test set that was held from the original dataset (see Figure 25). This test set has the same proportion of churners as the original unbalanced dataset. By using a highly unbalanced dataset, the results of the algorithms will be comparable to what the company would see if implemented into their standard business practices. To assess performance,

the confusion matrices, recall scores, and AUC-ROC will be compared; ANOVA tests will be used to compare the performance metrics amongst trials to determine which method is significantly better at predicting employee churn.

6.1 CONFUSION MATRIX

Confusion matrices show the classification of the algorithm compared to the actual class of the data objects. For this research model, a lower number of false negatives is preferred as that indicates a more conservative classification method. For interpretability, in Table 13 and Table 14, only the classifications for actual churners are displayed since this is the primary class of interest.

Table 13: Summary of Confusion Matrices from Models

			Predicted									
			0					1				
			k = 1	k = 3	k = 5	Ran.	None	k = 1	k = 3	k = 5	Ran.	None
Actual	1	Trial 1	DT	24	5	24	5	1	0	19	0	19	23
			SVM	17	17	17	2	17	7	7	7	22	7
			MLP	24	0	24	24	24	0	24	0	0	0
		Trial 2	DT	5	2	23	2	3	18	21	0	21	20
			SVM	1	19	19	19	0	22	4	4	4	23
			MLP	23	0	23	23	0	0	23	0	0	23
		Trial 3	DT	2	2	8	3	8	21	21	15	20	15
			SVM	12	12	12	12	12	11	11	11	11	11
			MLP	23	23	23	0	0	0	0	0	23	23
		Trial 4	DT	10	9	14	4	9	13	14	9	19	14
			SVM	2	16	16	2	16	21	7	7	21	7
			MLP	23	23	23	23	23	0	0	0	0	0
		Trial 5	DT	5	12	12	5	12	18	11	11	18	11
			SVM	18	18	18	18	18	5	5	5	5	5
			MLP	23	23	23	23	0	0	0	0	0	23
		Trial 6	DT	4	12	11	5	11	19	11	12	18	12
			SVM	1	1	18	1	18	22	22	5	22	5
			MLP	23	23	23	23	23	0	0	0	0	0
		Trial 7	DT	7	3	8	8	7	16	20	15	15	16
			SVM	1	16	1	16	16	22	7	22	7	7
			MLP	23	0	23	23	0	0	23	0	0	23
		Trial 8	DT	4	1	2	1	2	16	22	21	22	21
			SVM	14	14	14	14	14	9	9	9	9	9
			MLP	23	23	23	23	1	0	0	0	0	22
		Trial 9	DT	2	0	0	6	0	21	23	23	17	23
			SVM	1	1	11	1	11	22	22	12	22	12
			MLP	23	0	0	23	23	0	23	23	0	0
		Trial 10	DT	2	3	22	5	13	21	20	1	18	10
			SVM	2	15	15	2	15	21	8	8	21	8
			MLP	23	23	20	23	23	0	0	3	0	0

Table 14: False Negatives and True Positive Summary

	False Negative						True Positive									
	DT		SVM		MLP		DT		SVM		MLP		FN		TP	
	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD	Avg	SD
k = 1	6.5	6.6	6.9	7.4	23.1	0.3	16.3	6.29	16.2	7.22	0	0	12.17	9.61	10.83	9.44
k = 3	4.9	4.5	12.9	6.6	13.8	11.9	18.2	4.49	10.2	6.51	9.3	12.01	10.53	8.94	12.57	8.99
k = 5	12.4	8.5	14.1	5.3	20.5	7.3	10.7	8.31	9	5.25	2.6	7.23	15.67	7.74	7.43	7.67
Ran.	4.4	2	8.7	7.7	20.8	7.3	18.7	2	14.4	7.83	2.3	7.27	11.30	9.29	11.80	9.30
None	6.6	4.8	13.7	5.4	11.7	12.1	16.5	4.93	9.4	5.30	11.4	12.02	10.67	8.42	12.43	8.39
Avg	6.96		11.26		17.98		16.08		11.84		5.12					
SD	6.18		6.92		9.61		6.09		6.89		9.60					

Since the misclassification cost of an employee that actually leaves a company is higher than for one that ends up staying with the company, it is more important for the true positive rate to be high and the false negative rate to be low. The true negative and false positive are not the focus for this enterprise problem. Judging from only the confusion matrices, decision tree stood out as a better and more consistent classifier of churners. There wasn't, however, a single undersampling method that obviously increased true positive classification.

6.2 RECALL

Recall represents the number of data objects from the class of interest that were correctly classified by the model. This is an important metric for this research than accuracy because a higher score would be indicative of a better classifier for churners.

Table 15: Recall Scores for Models

		k = 1	k = 3	k = 5	Random	None
Trial 1	**DT**	0.	0.7917	0	0.7917	0.9583
	SVM	0.2917	0.2917	0.2917	0.9167	0.2917
	MLP	0	1	0	0	0
Trial 2	**DT**	0.7826	0.9130	0	0.9130	0.8696
	SVM	0.9565	0.1739	0.1739	0.1739	1
	MLP	0	1	0	0	1
Trial 3	**DT**	0.9130	0.9130	0.6522	0.8696	0.6522
	SVM	0.4783	0.4783	0.4783	0.4783	0.4783
	MLP	0	0	0	1	1
Trial 4	**DT**	0.5652	0.6087	0.3913	0.8261	0.6087
	SVM	0.9130	0.3043	0.3043	0.9130	0.3043
	MLP	0	0	0	0	0
Trial 5	**DT**	0.7826	0.4783	0.4783	0.7826	0.4783
	SVM	0.2174	0.2174	0.2174	0.2174	0.2174
	MLP	0	0	0	0	1
Trial 6	**DT**	0.8261	0.4783	0.5217	0.7826	0.5217
	SVM	0.9565	0.9565	0.2174	0.9565	0.2174
	MLP	0	0	0	0	0
Trial 7	**DT**	0.6957	0.8696	0.6522	0.6522	0.6957
	SVM	0.9565	0.3043	0.9565	0.3043	0.3043
	MLP	0	1	0	0	1
Trial 8	**DT**	0.8261	0.9565	0.9130	0.9565	0.9130
	SVM	0.3913	0.3913	0.3913	0.3913	0.3913
	MLP	0	0	0	0	0.9565
Trial 9	**DT**	0.9130	1	1	0.7391	1
	SVM	0.9565	0.9565	0.5217	0.9565	0.5217
	MLP	0	1	1	0	0
Trial 10	**DT**	0.9130	0.8696	0.0435	0.7826	0.4348
	SVM	0.9130	0.3478	0.3478	0.9130	0.3478
	MLP	0	0	0.1304	0	0

Table 16: Recall Summary

	k = 1		k = 3		k = 5		Ran		None			
	Avg	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**
DT	0.722	0.276	0.788	0.195	0.465	0.361	0.810	0.087	0.713	0.208	0.700	0.264
SVM	0.703	0.316	0.442	0.284	0.390	0.229	0.622	0.337	0.407	0.231	0.513	0.299
MLP	0.0	0.0	0.400	0.516	0.113	0.314	0.100	0.316	0.496	0.523	0.222	0.416
Avg	0.475		0.543		0.323		0.511		0.539			
SD	0.414		0.388		0.333		0.402		0.363			

Unsurprisingly, the models determined best from the confusion matrix also have very high recall scores, as shown in Table 15 and Table 16. Decision tree models continue to outperform other classification models. Undersampling of k = 3 have a slightly higher average recall scores compared to the other undersampling treatments, but all methods have a high variance in the results.

6.3 AUC-ROC

The ROC curve graphs the false positive rate to the true positive rate for a given model; the area under the ROC curves were calculated and compared to determine a superior model (see Table 17 and Table 18). The closer the AUC is to one, the better the algorithm at predicting the churning class.

Table 17: AUC-ROC for Models

		k = 1	k = 3	k = 5	Random	None
Trial 1	**DT**	0.8170	0.9901	0.8170	0.9043	0.9993
	SVM	0.9094	0.9096	0.9096	0.9307	0.9096
	MLP	0.4837	0.5000	0.5000	0.4869	0.6440
Trial 2	**DT**	0.9908	0.9982	0.8342	0.9540	0.9849
	SVM	0.9390	0.9390	0.9390	0.9390	0
	MLP	0.6389	0.4973	0.5000	0.4864	0.5000
Trial 3	**DT**	0.9722	0.9989	0.9467	0.9759	0.9467
	SVM	0.9112	0.9112	0.9112	0.9112	0.9112
	MLP	0.5187	0.5000	0.5000	0.5000	0.5027
Trial 4	**DT**	0.8789	0.9603	0.6875	0.9803	0.9603
	SVM	0.9327	0.9308	0.9308	0.9327	0.9308
	MLP	0.5163	0.5078	0.5000	0.5716	0.5000
Trial 5	**DT**	0.9240	0.7310	0.9217	0.9240	0.9368
	SVM	0.8856	0.8856	0.8855	0.8856	0.8856
	MLP	0.4783	0.4810	0.6342	0.4973	0.5027
Trial 6	**DT**	0.9972	0.7174	0.8726	0.8816	0.9187
	SVM	0.8802	0.8802	0.8143	0.8802	0.8143
	MLP	0.7591	0.5000	0.5000	0.4891	0.5938
Trial 7	**DT**	0.8466	0.9988	0.7962	0.8255	0.9282
	SVM	0.9197	0.9310	0.9197	0.9312	0.9312
	MLP	0.4701	0.5000	0.4918	0.5662	0.5082
Trial 8	**DT**	0.9200	0.9824	0.9534	0.9824	0.9734
	SVM	0.9048	0.9050	0.9048	0.9049	0.9048
	MLP	0.5378	0.5000	0.5000	0.4783	0.5877
Trial 9	**DT**	0.9969	0.9991	1	0.8977	1
	SVM	0.9421	0.9421	0.9475	0.9421	0.9475
	MLP	0.5512	0.5000	0.5000	0.4638	0.5000
Trial 10	**DT**	0.9722	0.9918	0.9341	0.9640	0.9597
	SVM	0.9282	0.8912	0.8912	0.9282	0.8912
	MLP	0.4563	0.5000	0.5652	0.4809	0.5000

Table 18: AUC-ROC Summary

	k = 1		k = 3		k = 5		Ran		None			
	Avg	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**	**Avg**	**SD**
DT	0.932	0.066	0.937	0.113	0.876	0.094	0.929	0.052	0.961	0.029	0.927	0.078
SVM	0.915	0.021	0.913	0.022	0.905	0.038	0.919	0.022	0.813	0.288	0.893	0.132
MLP	0.541	0.093	0.499	0.007	0.519	0.046	0.502	0.037	0.534	0.054	0.519	0.055
Avg	0.796		0.783		0.767		0.783		0.769			
SD	0.194		0.214		0.189		0.206		0.243			

Decision Tree models had significantly higher average AUC-ROC scores when compared to SVM and MLP models. Undersampling k = 1 have slightly higher average AUC-ROC when compared to the other undersampling methods, however the high variance for each method implies that the classification method could impact the undersampling's effectiveness. Further analysis shall be conducted to determine which classification algorithm and undersampling method performs best.

6.4 STATISTICAL ANALYSIS

In order to draw significant conclusions from this research, analysis of variance (ANOVA) tests was conducted on the effect of the classification method, undersampling, and the interaction of the two treatments had on the recall, AUC-ROC, true positive rate, and F-Measure for the models constructed in the study. The purpose of these tests is to conclude whether or not the undersampling or classification methods had a significant impact on improving the evaluation scores.

6.4.1 Recall Analysis

The ANOVA output with the response variable set to recall shows a significant difference between the treatments of the study (see Figure 26). Upon further investigation from the effects test (Figure 27), the effect of the classification method alone as well as the combination of undersampling and classification method had significant impacts on the recall score. The undersampling method independently did not have a significant effect on recall.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	9.172588	0.655185	6.8548
Error	135	12.903372	0.095581	Prob > F
C. Total	149	22.075961		<.0001*

Figure 26: ANOVA for Recall Scores

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Undersampling	4	4	0.9939611	2.5998	0.0389*
Method	2	2	5.7983839	30.3325	<.0001*
Undersampling*Method	8	8	2.3802435	3.1129	0.0029*

Figure 27: Effect Test for Recall Scores

In Figure 28, a Tukey All Pairwise Comparison was created to compare the three classification methods. As expected from the initial analysis, decision tree and SVM models had statistically significantly higher recall scores compared to the MLP models. Decision tree models also were statistically significant higher recall than SVM models. With respect to increasing recall score, there wasn't a single model that was significantly better than the other treatment combinations.

Level				Least Sq Mean
DT	A			0.69952899
SVM		B		0.51297101
MLP			C	0.22173913

Figure 28: Tukey Comparison for Recall Scores by Classification Methods

Level				Least Sq Mean
random,DT	A			0.80960145
k = 3,DT	A			0.78786232
k = 1,DT	A			0.72173913
None,DT	A			0.71322464
k = 1,SVM	A			0.70307971
random,SVM	A			0.62210145
None,MLP	A	B		0.49565217
k = 5,DT	A	B	C	0.46521739
k = 3,SVM	A	B	C	0.44221014
None,SVM	A	B	C	0.40742754
k = 3,MLP	A	B	C	0.40000000
k = 5,SVM	A	B	C	0.39003623
k = 5,MLP		B	C	0.11304348
random,MLP		B	C	0.10000000
k = 1,MLP			C	1.9429e-16

Figure 29: Tukey Comparison for Recall Scores by Undersampling x Classification Method

In Figure 29, a Tukey comparison of all combinations of undersampling and classification method was conducted. While there is not a single combination of treatment methods that outperforms the other models, it is worthy to note that the unbalanced MLP models have comparable performance to the decision tree and SVM models.

6.4.2 AUC-ROC Analysis

In the ANOVA test, there was a statistically significant difference between the twelve treatments and their AUC-ROC values (see Figure 30). The corresponding effect test shown in Figure 31 concluded that the classification method had an effect on the AUC-

ROC for a model, while the undersampling and interaction of undersampling and method did not significantly alter the evaluation metric.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	5.2574924	0.375535	43.3448
Error	135	1.1696277	0.008664	Prob > F
C. Total	149	6.4271202		<.0001*

Figure 30: ANOVA for AUC-ROC Scores

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Undersampling	4	4	0.0168141	0.4852	0.7466
Method	2	2	5.1236168	295.6874	<.0001*
Undersampling*Method	8	8	0.1170615	1.6889	0.1065

Figure 31: Effect Test for AUC-ROC Scores

Further analysis concluded that decision tree and SVM models had higher AUC-ROC scores than MLP (see Figure 32). Since there wasn't a significant effect of the interaction of undersampling and classification method, a Tukey comparison was not necessary to conclude that all decision tree and SVM models had statistically significant higher AUC-ROC scores compared to all MLP models.

Level			Least Sq Mean
DT	A		0.92690112
SVM	A		0.89288272
MLP		B	0.51894363

Figure 32: Tukey Comparison for AUC-ROC Scores by Classification Methods

6.4.3 F-Measure Analysis

An ANOVA test for the response F-Measure was conducted and was found to have had a significant increase for F-measure with respect to undersampling, classification method, and undersampling and classification method combination, as shown in Figure 33 and Figure 34. The Tukey comparison test in Figure 35 confirms that decision tree models had statistically significant higher F-measure scores than SVM or MLP models, identical to the conclusion from the recall analysis.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	14	12.997586	0.928399	35.6768
Error	135	3.513035	0.026022	Prob > F
C. Total	149	16.510620		<.0001*

Figure 33: ANOVA for F-Measure Score

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Undersampling	4	4	0.342826	3.2936	0.0130*
Method	2	2	12.066676	231.8510	<.0001*
Undersampling*Method	8	8	0.588084	2.8249	0.0063*

Figure 34: Effect Test for F-Measure Score

Level				Least Sq Mean
DT	A			0.73142733
SVM		B		0.50290235
MLP			C	0.04898131

Figure 35: Tukey Comparison for F-Measure by Classification Method

Unlike the analysis from the other evaluation metrics, undersampling had a significant effect on increasing F-Measure in this experiment. Figure 36 shows the Tukey comparison test for undersampling; k = 5 undersampling significantly underperformed when compared to the other undersampling methods. This is surprising as the control dataset that didn't address the imbalance performed significantly better than compared to the k = 5 datasets.

Level			Least Sq Mean
k = 3	A		0.46028287
None	A		0.45711122
random	A		0.45333269
k = 1	A	B	0.43423908
k = 5		B	0.33388578

Figure 36: Tukey Comparison for F-Measure by Undersampling Method

In Figure 37, the Tukey all pairwise comparison show that all MLP models had a significantly lower F-Measure score compared to other treatment combinations. In general, decision tree and better-balanced models had a higher performance. Decision tree models with random, k = 1, and k = 3 undersampling as well as the original unbalanced dataset performed significantly better than undersampling of k = 5 for all classification methods.

Level						Least Sq Mean
random,DT	A					0.84006641
k = 3,DT	A					0.82413909
k = 1,DT	A	B				0.77269665
None,DT	A	B	C			0.73495508
None,SVM		B	C	D		0.53523862
k = 1,SVM		B	C	D		0.53002060
random,SVM			C	D		0.49993166
k = 5,DT				D		0.48527942
k = 3,SVM				D		0.47601986
k = 5,SVM				D		0.47330102
None,MLP					E	0.10113997
k = 3,MLP					E	0.08068966
k = 5,MLP					E	0.04307692
random,MLP					E	0.02000000
k = 1,MLP					E	8.6736e-16

Figure 37: Tukey Comparison for F-Measure by Undersampling and Classification Method

6.5 BEST MODEL CONCLUSIONS

Based on the resulting evaluation metrics and statistical analysis conducted, decision tree is the superior classifier for employee churn. Although a single undersampling method did not prove to have a significantly better classification performance, k = 3 undersampling did outperform other methods in nearly every evaluation metric for decision tree models. Therefore, undersampling k = 3 using decision tree classification is the superior predictor of employee churn. It is worth noting that although k = 3 was successful in aiding in classification performance, decision tree classification proved to be nearly as successful with the original unbalanced dataset.

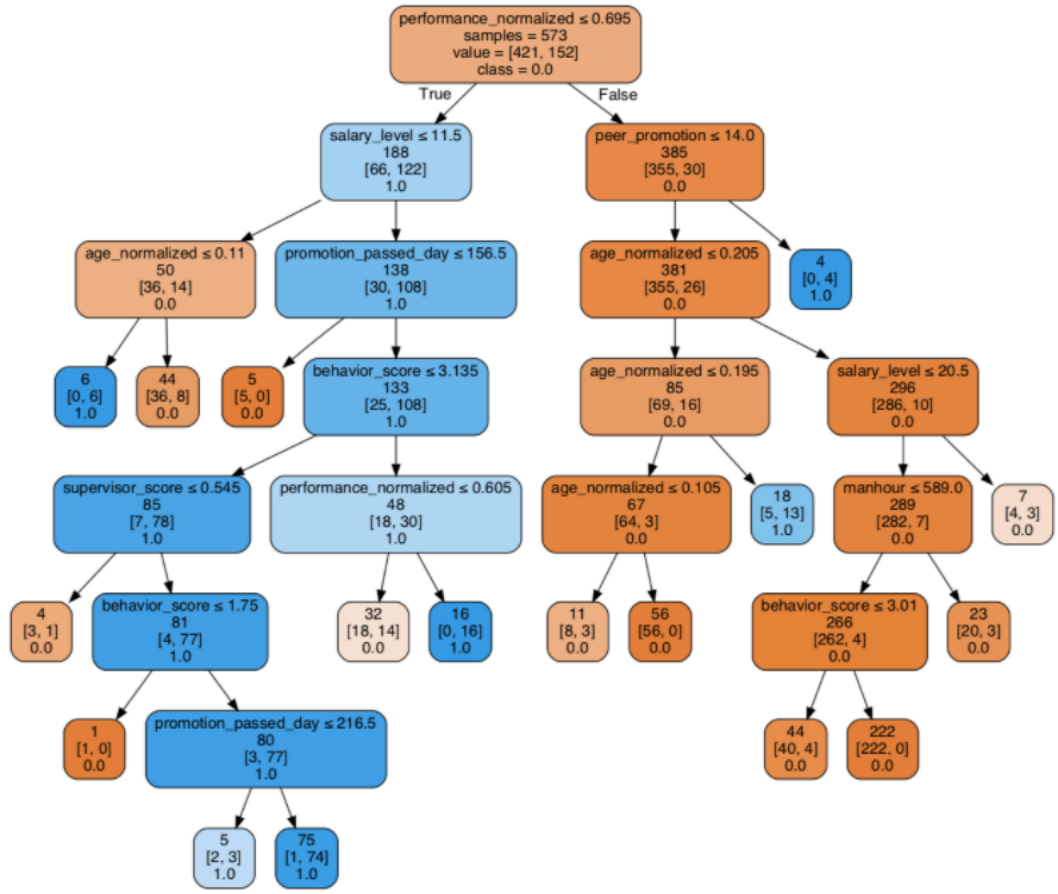

Figure 38: Decision Tree Model for Classifying Employee Churn

Table 19: Evaluation Metrics Summary for Best Model

Model	Recall	Accuracy	F-Measure	TP/FN
k = 3 x DT	0.7143	0.9304	0.6962	34/22

The final decision tree model with undersampling k = 3 is shown above in Figure 38; the corresponding evaluation metrics can be seen in Table 19. The resulting decision tree uses employee data from performance score normalized, salary level, number of peer promotions, age normalized, days since last promotion, behavior score, supervisor score, and training manhours to classify the data. Most notably, the initial node splits the data by a threshold performance score. From the right side of the tree, if an employee is a high performer (greater than the threshold of 0.695) and their peers have been promoted many

times (greater than 14), then they are classified as churners. This information can be used by the company to monitor their employees for similar trends in order to mitigate their churn risk. A way to prevent these employees from churning would be to have their own success addressed as often as their peers, when applicable. From the left side of the tree, underperforming employees with a lower salary level and younger compared to the rest of the company have been classified as churners. These traits could potentially be indicators of a new employee overwhelmed or undertrained in their position, leading to them leaving out of frustration. Knowing these are indicators of churn, the company can identify those employees that are struggling to adjust to the working dynamic and offer additional training or mentors for their position to ease the transition. Interpreting and applying the resulting decision rules and analyzing the contributors to churn had the potential of increasing employee retention due to early detection and involvement by the company.

7. CONCLUSIONS

7.1 SUMMARY

In this research, different combinations of undersampling and classification algorithms are tested together in hopes of finding an optimal combination for employee churn prediction. Thorough data cleaning preprocessing was applied to the dataset in order to address missing values, outliers, and the dataset dimensions in a way that wouldn't hinder the algorithm performance. Experimentation concluded that decision tree is consistently more successful in identifying churners than SVM or MLP across all evaluation metrics. The best model conclusion coupled decision tree classification with $k = 3$ undersampling. Decision rules can be extracted from the resulting algorithms in order to improve business processes and standard practices to help increase employee retention. The significant attributes discovered in this study can be used by the company to pinpoint common trends in their churning employees and focus on reversing these effects. Similar analysis can be conducted by other companies concerned with their churn rates to identify churner's trends in their business in hopes of mitigating the risk in time to retain their employees.

7.2 LIMITATIONS

A limitation to the employee churn classification problem is the privacy of the company's employees. Employee churn problems generally use a company's human resource department data from the time of employee hiring. If new data is collected from the employees, there needs to be full transparency in how the personal data will be used. By

being fully transparent, bias could be introduced into the data and effect the resulting models.

7.3 ETHICAL CONSIDERATIONS

When using data mining to solve enterprise problems, the ethical ramifications must be taken into consideration before adopting any new practices. With personal data, there is always a concern with incorporating societal biases into the algorithms. Any biased assumptions that are followed in society by human decision makers will be portrayed as patterns in the data; this pattern may not be attributed to the data object but rather the continued bias of humans. Creating a data mining algorithm is often mistaken as a way to solve this bias, but by using the historical biased data, the invalid assumptions are modeled into the algorithm and the bias continues. In the case of employee churn, the ramifications of bias pose a risk of targeting a certain subset of the employee pool for retention where it may not be needed.

Another unchecked assumption with employee churn is data object sameness. Assuming that all employees are in the same situation with the same preferences in regard to their employment would be incorrect; additionally, assuming that all employees hold the same value to the company would also be false. A CEO's churn will have a much bigger impact on the company's system than an administrative person's churn. In this research, data classification imbalance was addressed through the evaluation metrics used to portray success, but data object dissimilarity was not taken into account for classification. Lastly, the limits of modeling employee churn through the data collected must also be taken into account with the algorithm results. The assumption that an employee's

likelihood to churn can be predicted by the included attributes will always leave out other reasons for churn that can't be portrayed with data. The models will only assimilate reality, but not be a perfect representation. Therefore, by treating the model outcomes as fact and only treating those classified as churn risks, the company would be alienating non-churn employees and risk increasing their likelihood of churn due to feeling undervalued. Addressing churn at a company-wide scale as well as an individual basis is necessary to keep the balance.

7.4 FUTURE WORK

Future work could build off this research by including a predictive model of employee value to the company. Employee value could be used after classification to help prioritize which employees to target for churn risk mitigation, addressing the more valuable employees first as they would be the most challenging and costly to replace. Having access to their inherent value will further help narrow down the data object list into a manageable subset that can be addressed by the company on a more personal level. Another aspect to expand on with this research would be its robustness when applied within different industries. There is a vast amount of literature analyzing customer churn within the telecommunication industry, and a great deal of employee churn research has also been made in this field due to availability of data. If the findings from this thesis as well as other pieces of research in employee churn were applied in the defense, e-commerce, or biomedical industry, would the success in predicting employee churn transfer to these other industries? Or are there different relationships and patterns within other industries that is not apparent in telecommunications? As machine learning and data

mining techniques are becoming more widely accessible to all levels of business, big data will become more available for advanced analytics, such as employee churn prediction. Lastly, future research could look into the impact of oversampling on churn prediction. Past research has often concluded that undersampling will aide with classification in unbalanced datasets; however, with this research undersampling had no significant effect on churn classification. Understanding when undersampling or oversampling can assist in machine learning would improve performance and applications for many enterprise problems.

BIBLIOGRAPHY

Alamsyah, A., & Salma, N. (2018). A Comparative Study of Employee Churn Prediction Model. *2018 4th International Conference on Science and Technology (ICST)*, 1–4. https://doi.org/10.1109/ICSTC.2018.8528586

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254. https://doi.org/10.1016/j.neucom.2016.12.009

Chang, H.-Y. (2009). *Employee Turnover: A Novel Prediction Solution with Effective Feature Selection*. 5.

D, A., & B, A. A. (2013). *Analyzing Employee Attrition Using Decision Tree Algorithms*.

Dogan, N., & Tanrikulu, Z. (2013). A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management*, *14*(2), 105–124. https://doi.org/10.1007/s10799-012-0135-8

Dutta, S., & Bandyopadhyay, S. (2020). Employee Attrition Prediction Using Neural Network Cross Validation Method. *International Journal of Commerce and Management Research.*

Fletcher, T. (n.d.). *Support Vector Machines Explained*. 19.

Hong, W.-C., & Chao, R.-M. (2007). A Comparative Test of Two Employee Turnover Predicting Models. *International Journal of Management.* Retrieved August 12, 2020, from https://eservice.oit.edu.tw/fund/96/file/96_4_0152.pdf

Jafari, R. (2020). *Bayesian Classification* [PowerPoint presentation]. California Polytechnic State University San Luis Obispo, IME 372.

Jafari, R. (2020). *Classification* [PowerPoint presentation]. California Polytechnic State University San Luis Obispo, IME 372.

Jafari, R. (2020). *Decision Trees* [PowerPoint presentation]. California Polytechnic State University San Luis Obispo, IME 372.

Jafari, R. (2020). *MLP Classification* [PowerPoint presentation]. California Polytechnic State University San Luis Obispo, IME 372.

Jain, D. (2017). *Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods*. 21.

Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). *Towards applying Data Mining Techniques for Talent Mangement*. 6.

Lazarov, V., München, T. U., Capota, M., & München, T. U. (2007). Churn Prediction. *Business Analytics Course. TUM Computer Science*.

Ma, X., Zhai, S., Fu, Y., Lee, L. Y., & Shen, J. (2019). *Predicting the Occurrence and Causes of Employee Turnover with Machine Learning*. *8*(3), 11.

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence*, *5*(9). https://doi.org/10.14569/IJARAI.2016.050904

Richter, Y., Yom-Tov, E., & Slonim, N. (2010). Predicting customer churn in mobile networks through analysis of social groups. *Proceedings of the 2010 SIAM International Conference on Data Mining*, 732–741. https://doi.org/10.1137/1.9781611972801.64

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006. https://doi.org/10.1016/j.eswa.2010.07.134

Shankar, R. S., Rajanikanth, J., Sivaramaraju, V. V., & Murthy, K. V. S. S. R. (2018). PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING. *2018 Ieee International Conference on System, Computation, Automation and Networking (Icscan)*, 1–8. https://doi.org/10.1109/ICSCAN.2018.8541242

Sundarkumar, G. G., & Ravi, V. (2015). A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance. *Engineering Applications of Artificial Intelligence*, *37*, 368–377. https://doi.org/10.1016/j.engappai.2014.09.019

Tsai, C.-F., & Chen, M.-Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications*, *37*(3), 2006–2015. https://doi.org/10.1016/j.eswa.2009.06.076

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). *A comparison of machine learning techniques for customer churn prediction | Elsevier Enhanced Reader*. https://doi.org/10.1016/j.simpat.2015.03.003

Yigit, I. O., & Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–4. https://doi.org/10.1109/IDAP.2017.8090324

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2019). Employee Turnover Prediction with Machine Learning: A Reliable Approach. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Intelligent Systems and Applications* (Vol. 869, pp. 737–758). Springer International Publishing. https://doi.org/10.1007/978-3-030-01057-7_56