NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

# Multiclass Classification of Motor Insurance Customers in Portugal

Ekaterina Mylnikova

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# MULTICLASS CLASSIFICATION OF MOTOR INSURANCE CUSTOMERS IN PORTUGAL

by

Ekaterina Mylnikova

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

**Advisor:** Mauro Castelli

July 2021

# DEDICATION

This report is dedicated to my family for their endless support and unconditional love.

# ACKNOWLEDGMENTS

I would like to thank my supervisor Professor Mauro Castelli for all the mentoring, assistance, patience, and understanding.

I wish to express my gratitude to Magdalena Neat for the opportunity, her time, and dedication. I would also like to thank Franklin Minang for his guidance and support during the project.

To the Data Mining team and my colleagues during this internship.

To my family who believed in me and supported my decision to move far away from home to get the degree.

To my beloved friends who supported me during my studying, internship, and writing the report.

# ABSTRACT

The insurance market is highly competitive. To stay in line with other companies in today's world, it is not enough for a company to have the best price. The most important move now is to make a personalized offer to each client. Insurance companies have an enormous amount of data that can be used to understand their customers better. What do they want? What offer would attract new clients, and what offer would keep existing customers from leaving? The project aims to classify customers' profiles based on their individual preferences in motor insurance.

# KEYWORDS

Data Mining; Machine Learning; Motor Insurance; Classification; Decision Tree; Artificial Neural Network; Random Forest; Gradient Boosting

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

**Abbreviations of Coverages**

| | |
|---|---|
| **RCV** | Third party liability |
| **ASB/AST** | Travel Assistance |
| **PTJ** | Legal Protection |
| **ODV** | Vehicle Occupants |
| **QIV** | Isolated Breakage of Glass |
| **CCC/CRO** | Crash, Collision, Roll-Over |
| **IRE** | Fire, Lightning Strike, Explosion |
| **FRB** | Theft or Robbery |
| **FDN** | Natural Phenomena |
| **ADV** | Acts of Vandalism |
| **VDS** | Replacement Vehicle - Accident |
| **VAV** | Replacement Vehicle - Breakdown |
| **BAG** | Baggage |
| **PMU** | Pack Relax |
| **PSR** | Pack Vintage |
| **PPE** | Pack Prestige |

# 1. INTRODUCTION

This report is a result of a 9-months internship at Grupo Ageas Portugal. The project took place from September 2019 to June 2020.

Grupo Ageas Portugal is one of the biggest and well-known insurance companies in Portugal.

Initially, the internship project was about creating a customer classification for a joint project of Ageas and Millennium BCP. The first month of work was dedicated to meetings with relevant stakeholders and gathering information about the new application. However, after a few weeks, it became clear that the launching of the application was postponed for an undefined period.

After analyzing projects in the company and doing research about them, a new project was defined.

Ageas offers different insurance products. This project is focused on motor insurance.

Motor Insurance is a kind of insurance policy that covers vehicles from potential risks. In Portugal, motor insurance is mandatory. Therefore, whenever a person buys any car, vehicle insurance is equally valuable and essential.

The biggest challenge for insurance companies is retaining existing customers and getting car owners to obtain motor insurance from their company. A customer can easily switch from one insurer to another. The competition will be won by the one who has a personalized offer with the best price for each client.

Considering the business needs and available data, the main objective of this project was to make classification of motor insurance customers profiles based on their preferences in packages and combination of coverages.

This report is divided into two parts. The first part provides a Literature Review of classification algorithms studied during the project and other topics used to obtain the results. The second part contains a description of the process used to build a classification model for car insurance clients.

# 2. LITERATURE REVIEW

## 2.1. DATA MINING PROCESS

In order to perform data mining analysis, there are standard data mining processes that can be followed during the project. Two most popular ones are CRISP-DM and SEMMA.

### 2.1.1. CRISP-DM

A cross-industry standard process for data mining known as CRISP-DM is an analytics model that describes the data science lifecycle. (Chapman, et al., 1999) The first version of this model, CRISP-DM 1.0, was published in 1999.

This process contains six steps that form a cycle.

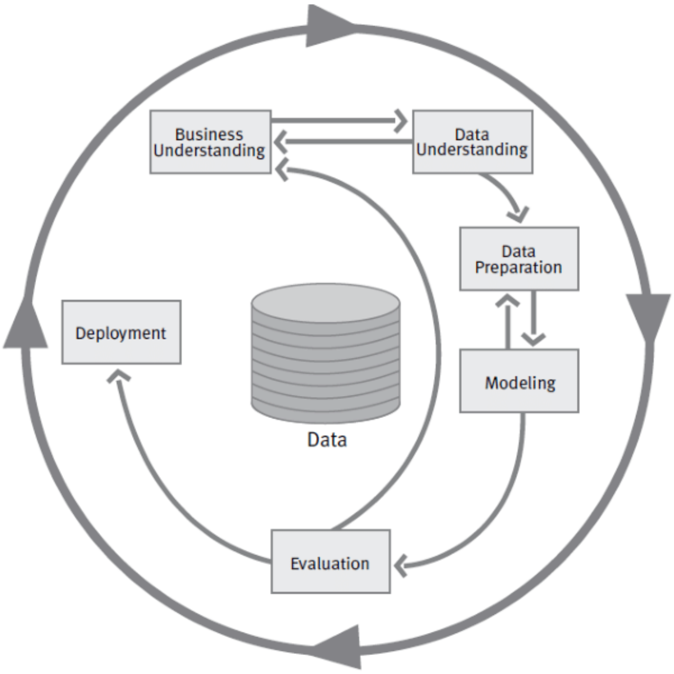


Figure 2.1 - CRISP-DM1 Process

1. Business understanding. The first phase is focused on understanding the project from a business perspective. Based on this, the data mining problem is defined, and a project plan is created.
2. Data understanding. This phase starts with data gathering and continues with data exploration. During this step, data findings can be discovered, and some data quality problems can be understood. This information is a basis for the next step.
3. Data preparation. This phase covers all data manipulations and changes needed to build the final dataset. It includes data cleaning, variable transformation, variable selection, creation of new variables.

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999, 2000). CRISP-DM 1.0.

4. Modeling. In this step, suitable machine learning models are tested as well as their parameters.
5. Evaluation. During this phase, the model chosen in the previous step is being tested, and the results are analyzed. If the model shows good results, a business decision is made either to use this model or not at the end of this step.
6. Deployment. This phase is what differentiates the academic and business projects. In business, model creation is not the last step. The model is generally useless if it cannot be deployed. During this phase, various teams in the company work together to bring the model to life.

### 2.1.2. SEMMA

The SEMMA model is specific to SAS. According to SAS Institute, data mining is the process of Sampling, Exploring, Modifying, Modeling, and Assessing large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage (SAS Institute, 2017).

This model contains all the steps needed to perform a data mining process in any industry.

1. Sampling. Splitting the data into train, validation, and test datasets.
2. Exploring. Data exploration to get the main findings, discover trends, anomalies, and relationships.
3. Modifying. Creation, selection, and transformation of variables to change the dataset.
4. Modeling. Applying Machine Learning models to find a combination of data that reliably predicts a desired outcome.
5. Assessing. Evaluation of the results of the finding from the data mining process.

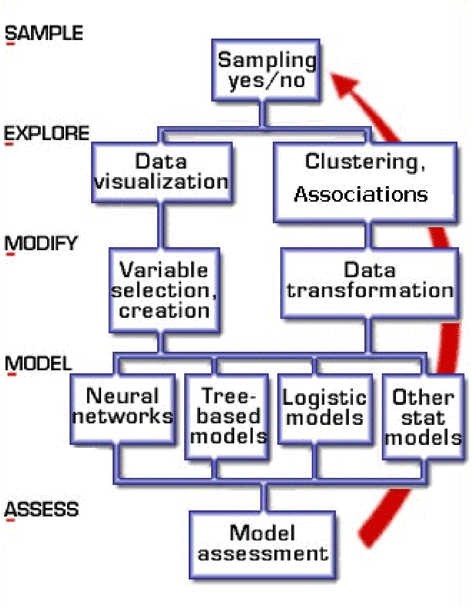


Figure 2.2 - SEMMA Process

## 2.2. TOOLS AND TECHNOLOGY

SAS is software developed by SAS Institute. This tool is used widely for data management, data mining, business intelligence, and machine learning.

There are various tools that SAS Institute offers for data analytics. In this project, SAS Enterprise Guide and SAS Enterprise Miner are two tools that were used on each step.

### 2.2.1. SAS Enterprise Guide

SAS Enterprise Guide is a point-and-click, menu- and wizard-driven tool that helps users analyze data and publish their results (SAS Institute, 2017). This tool allows users to manage and join multiple datasets to perform data analysis.

SAS EG makes it easy to join multiple data sources and visualize this process providing an intuitive process flow.

The most significant advantage of SAS Enterprise Guide is that it allows users to use code or a “point and click” interface to manipulate data.

### 2.2.2. SAS Enterprise Miner

SAS Enterprise Miner is one of the most common data mining software tools (Nisbet, Elder & Miner, 2009). It is a tool for analytical professionals who analyze big volumes of data and apply advanced analytics techniques.

SAS EM visualizes the process of data mining through the five-step SAS SEMMA model, described in the previous chapter. Users select the needed tab from the toolbar to build a process flow and drag and drop them, connecting with nodes. It supports several algorithms, such as neural networks, regression, decision trees, and others.

One of the advantages of SAS EM is that this tool allows users to understand key relationships and develop models clearly and quickly by using an intuitive interface.

SAS Enterprise Miner is a universal tool and can be used in different industries.

## 2.3. DATA MINING TECHNIQUES

Data mining is a process of discovering patterns and findings in data sets. It is also known as knowledge discovery in databases (Han, 1996). The knowledge extracted from historical data helps users make predictions. These techniques are widely used in business to make data-driven decisions.

Data mining techniques were established as the result of a research and product development process (Rygielski et al., 2002). Data mining began with the initial storage of data on computers and has progressed with improvements in data access to the point where users can now navigate through data in real-time.

Some of the most used data mining techniques are classification, clustering, regression, association rules, outlier detection, prediction, visualization (Ramageri, 2010).

The ones that were used in this project are:

- Data cleaning and preparation. In the data mining process, cleaning and preparing data is crucial. Raw data must be cleansed and structured before it can be used in various analytic ways. Several parts of data modeling, transformation, data migration, and aggregation are applied in data cleaning and preparation. Recognizing the basic features and attributes of data is crucial in identifying the optimal use of data.
- Tracking patterns. Pattern recognition is the following data mining approach. It entails recognizing and tracking data patterns to make conclusions regarding business outcomes.
- Outlier detection. It helps to detect anomalies in the dataset, treat them and improve the results of future work with this data.
- Classification. Classification data mining techniques are used to examine the data attributes from various sources. After determining the key properties of these data kinds, they can be categorized or classified. It is used to understand data and separate it into different classes for future use.
- Prediction. It is one of the essential features of data mining. Prediction works by analyzing the trends from the historical data and extending them into the future. Thus, it gives insights about patterns that can appear in the future.
- Visualization. Data visualization is one of the underestimated data mining techniques. It provides insights from the data based on visual impressions that people can see. Nowadays, data visualization can be interactive and more descriptive than simple numbers in the table.

## 2.4. Classification Problem

Classification problems occur in different aspects of life and have two meanings (Michie et al., 1994). First, a set of observations may be given to identify classes or clusters in the data. Second, there already can be defined classes, and the goal is to describe a rule that explains existing classification. In machine learning, the first one is known as unsupervised learning, and the second is supervised learning.

In this project, the multiclass classification problem was solved by means of supervised learning.

As explained previously, in machine learning, classification is a problem of defining to which class a new observation belongs. When there are more than two classes to identify, the problem is called multiclass classification. Normally, one observation is assumed to belong to one class label, and the class labels are independent (Silva-Palacios, et al., 2017).

## 2.5. Machine Learning Classification Models

In this project, supervised learning was used to solve the multiclass classification problem. Unlike unsupervised learning, where the target is unknown, data has a predefined target variable in supervised learning.

A wide range of algorithms is available for a classification problem, each with its advantages and disadvantages. Unfortunately, none of the algorithms can solve all the challenges of supervised learning.

For this project, a business decision was made to use artificial neural networks, decision trees, gradient boosting, and random forest algorithms.

### 2.5.1. Artificial Neural Network

Artificial Neural Network is a machine learning algorithm created based on the knowledge of the brain: neurons are connected to other neurons, and it creates a network (Mitchel, 1997).

ANN is a massively parallel combination of basic processing units that can learn from the environment and store the learning in its connections (Haykin, 1999).

This algorithm is called a black box method since studying its structure will not give any understanding of the insights about the results.

Figure 2.3 illustrates a simple artificial neural network. Before building a neural network, the network topology has to be defined. The topology of the network describes the number of neurons and layers in the model.

Each layer of NN consists of neurons. The first layer is an input layer. The next layer is a hidden layer, and the last layer is the output layer. The connections between layers are called weights, and each neuron's value is multiplied by the value of each weight. The value of the connections is the only value that can be changed during the learning step.

To transform an input signal from each neuron, an activation function is applied. It combines inputs into a single output to be used by the next layer.

The Neural Network algorithm can work with incomplete data and, after finishing the training, can make predictions quite fast. But, on the other hand, it is a black box algorithm which means that it is difficult to interpret the results; and this algorithm can be time-consuming during the training step.

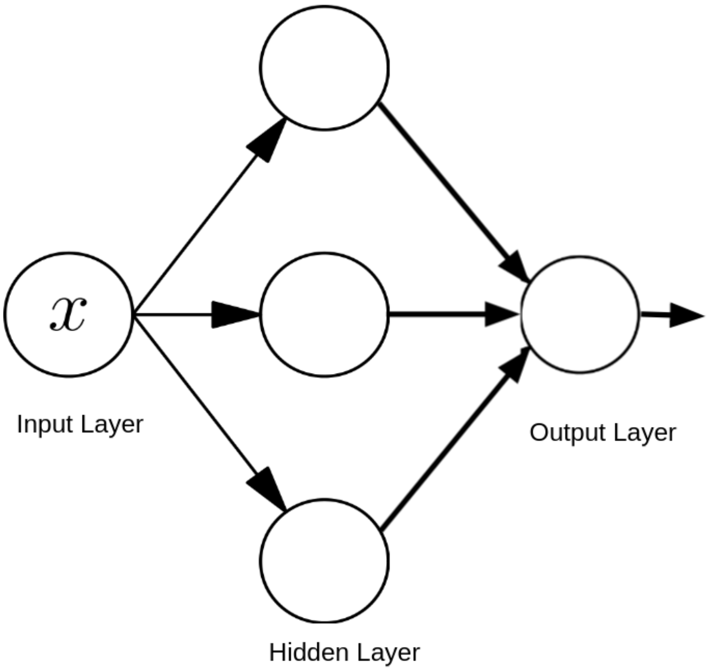


Figure 2.3 - Representation of Artificial Neural Network

### 2.5.2. Decision trees

Decision trees are one of the most popular algorithms due to their applicability in different fields.

The decision tree classifier is considered a multistage decision making (Safavian, & Landgrebe, 1991). The idea behind a multistage approach is breaking a big complex decision into a combination of several simple decisions; and this way to get the final solution.

This series of simple decision rules that a decision trees model produces can be easily understood and interpreted by people since it can be expressed in words (Berry & Linoff, 2009).

Decision trees models start from the root node and go to leaf nodes, classifying data. In between, there are decision nodes. Decision nodes have branches and are used to make a decision. Leaf nodes are the output of the decision. A decision tree asks a question and splits this tree into subtrees based on the answer.

A decision trees model is always shown upside down and starts with its root, the first decision node.

This algorithm doesn’t require treating missing values or feature selection on the data preparation step. It has a clear and intuitive visualization and can be easily explained. Decision trees can also be easily explained in words that make them suitable for various business cases.

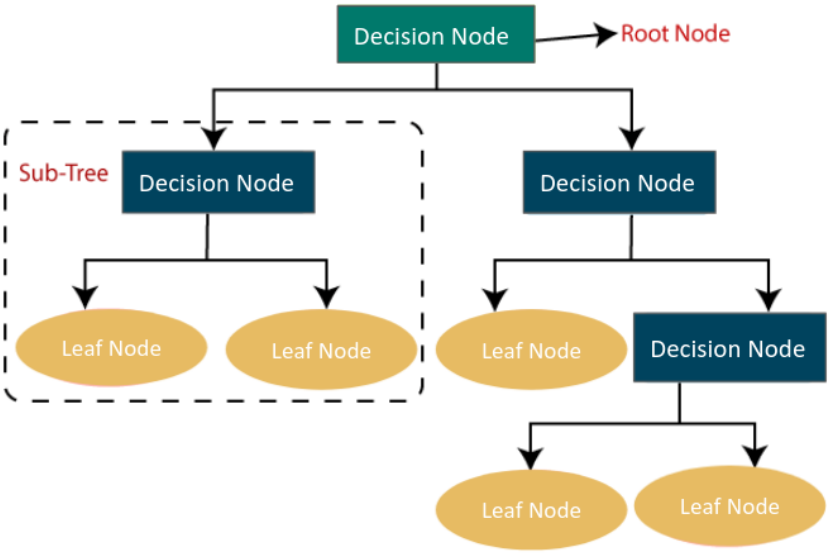


Figure 2.4 - An example of a decision tree

### 2.5.3. Random Forest

Random Forest is based on ensemble learning. It contains classification or regression trees (Breiman et al., 1984). These trees are trained on datasets created by random resampling on the initial dataset. In addition to this, this algorithm also chooses a random selection of features (Nagpal, 2017).

The algorithm takes the results from each tree, and, based on the votes of predictions, it predicts the final output. This technique leads to higher accuracy than a simple decision trees model and prevents overfitting.

Figure 2.5 illustrates a random forest algorithm process.

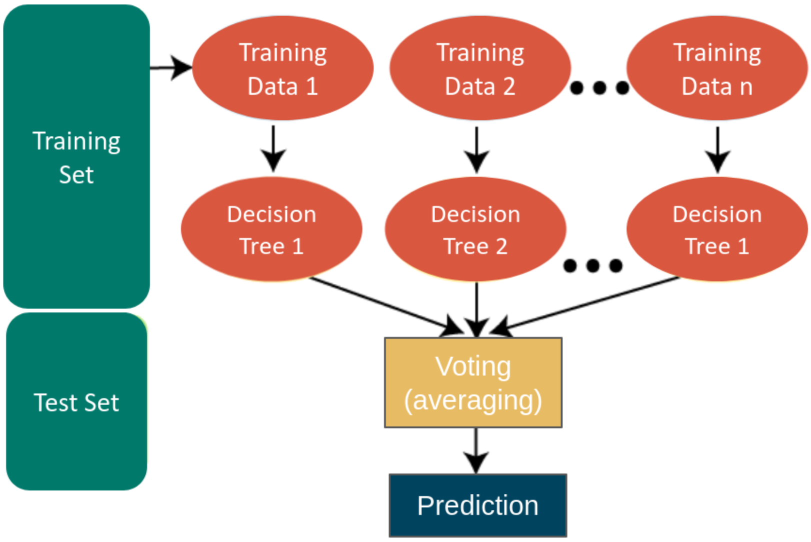


Figure 2.5 - Representation of a random forest

The same as the decision trees algorithm, random forest automatically applies feature selection. This algorithm is highly accurate and unlikely to overfit since it is based on the average results of individual decision trees. It can be used for both regression and classification problems. On the other hand, random forest is computationally more expensive, and the interpretation of the results is not as easy as a simple decision tree.

### 2.5.4. Gradient Boosting

The ensemble algorithms like random forests rely on simple averaging of models in them. The boosting method is based on a different strategy of ensembling. The idea behind boosting is to add a new weak basic model at each iteration and to train it considering all the errors in the model learned so far (Natekin &Knoll, 2013). The Gradient Boosting algorithm can be used for both classification and regression problems.

The most used implementation of gradient boosting is using decision trees (Dorogush et al., 2018).

Gradient Boosting algorithm requires minimum data preprocessing, same as decision trees. And considering its implementation, it provides a more accurate prediction. On the other hand, this algorithm is very computationally expensive, and it tends to overfit. Compared to a single decision tree, it is much harder to interpret a gradient boosting model results.

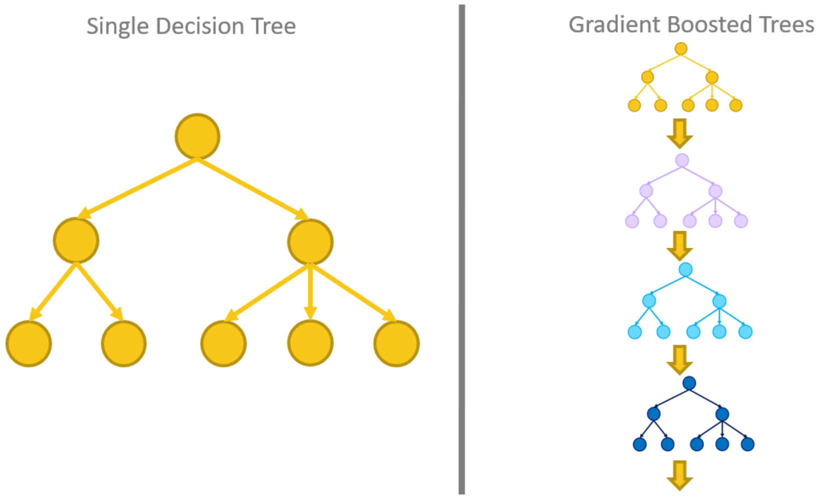


Figure 2.6 - Representation of Single Decision Tree VS Gradient Boosted Trees (Silipo, 2020)

## 2.6. MODEL PERFORMANCE EVALUATION

The process of evaluating the performance of a machine learning algorithm is important to select the model that shows the best results.

As a performance measure in this project, the F1 Score was used. The reason for this is an imbalanced dataset.

A confusion matrix is a table used to describe the performance of a classification model (Goutte, C., Gaussier, E., 2005).

| | Predicted class | | |
|---|---|---|---|
| Actual Class | | **Class = Positive** | **Class = Negative** |
| | **Class = Positive** | True Positive | False Negative |
| | **Class = Negative** | False Positive | True Negative |

Table 2.1 - Confusion Matrix

- True Positive. Correctly predicted positive values. It means that the actual class and the predicted class are the same.
- True Negative. Correctly predicted negative values.

- False Positive. Incorrect prediction. It occurs when the actual class is NEGATIVE, and the predicted class is POSITIVE.
- False Negative. Incorrect prediction. When the actual class is POSITIVE, and the predicted class is NEGATIVE.

Precision shows what proportion of predicted POSITIVE is actually POSITIVE.

Precision = TP ÷ (TP + FP)

Recall shows what proportion of actual POSITIVE is correctly classified.

Recall = TP ÷ (TP + FN)

F1 Score = 2 × (Precision × Recall) ÷ (Precision + Recall)

To calculate F1 Score for a multiclass classification problem, it is needed to take the average from the F1 Score for all classes. But in the case of an imbalanced dataset, it is recommended to use Weighted F1 Score.

To calculate Weighted F1, multiply the F1 Score by the number of observations in each class, sum them up, and divide by the total number of observations in the dataset.

The closer F1 Score to 1, the better is the model performance.

## 3. METHODOLOGY

### 3.1. Overview

Several different methodologies are available to use for a data mining project.

For this business case, CRISP-DM seemed like a more suitable methodology. The reason for that is that it contains the "Business understanding" step, a crucial step in a data mining project.

Therefore, CRISP-DM methodology was chosen to use with a few adjustments.

| | Sep | Oct | Nov | Dec | Jan | Feb | Mar | Apr | May | Jun |
|---|---|---|---|---|---|---|---|---|---|---|
| Onboarding Challenge | 🟩 | | | | | | | | | |
| | ▨ | | | | | | | | | |
| Business Understanding & Kick-off Presentation | | 🟥 | 🟥 | | | | | | | |
| | | ▨ | ▨ | | | | | | | |
| Data Exploration and Preparation | | | | 🟩 | 🟩 | 🟩 | | | | |
| | | | | ▨ | ▨ | | | | | |
| Modeling | | | | | | | 🟥 | 🟥 | 🟥 | |
| | | | | | | ▨ | ▨ | | | |
| Presenting Findings & Deploying Model | | | | | | | | | 🟩 | 🟩 |
| | | | | | | | | ▨ | | |
| Setup performance monitoring and define recommendations | | | | | | | | | | 🟥 |
| | | | | | | | | | ▨ | ▨ |
| Extending knowledge to future projects | | | | | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |
| | | | | | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |

Table 3.1 - Project Timeline. The color shows actual time; Grey shows initially planned time

Table 3.1 shows the roadmap of the project with a brief description based on the CRISP-DM methodology. Initially, the timeline was different, but some steps took more time than was planned as going on with the project. For example, the most time-consuming phase was data exploration and preparation, and the Modeling phase coincided with the beginning of Covid-19 and took longer than expected.

## 3.2. Business understanding

The first step of CRISP-DM methodology is business understanding.

Due to the complexity of motor insurance, it was essential to understand how this business works and the needs of the relevant stakeholders.

After onboarding meetings, one of the most critical problems was identified. This car insurance is sold through managers at the bank, so what would be the best-personalized combination of coverages that they can offer to each client?

During the Business understanding phase, the project objective and scope were defined.

**Objective:**

- Make a classification of car insurance customers' profiles based on their preferences in packages and combination of coverages using an advanced machine learning classification model.

**Scope:**

- Identify target classes based on combinations of insurance packages and coverages.
- Identify important variables through analyzing existing and creating new features.
- Apply different classification models and choose one with the best performance.
- Describe rules and patterns for each class.
- Deploy the model.

It was also important to understand what tools are used in the company. The most used tools are SAS Enterprise Guide and SAS Enterprise Miner. These two tools were chosen for this project since all the needed data is stored in the database connected to them.

During this phase, some potential business obstacles were identified.

**Challenges:**

- Car insurance is obligatory to have. Hence does it mean that customers buy the cheapest option with mandatory coverages?
- The product's price is higher than competitors'. Is there a significant variety of combination of packages and coverages, or do customers prefer to buy just the cheapest one?

## 3.3. DATA UNDERSTANDING

Since classification modeling had not been performed in the company before, it was necessary to understand available data sources and variables.

### 3.3.1. Data Sources

The main challenge of this part was that not all the relevant data was available to use for modeling. The reason for this is the relationship between the insurance company and the partner bank. Due to GDPR, the bank was no longer able to share the data with the company even though these clients were the company's clients.

Given these circumstances, there were 487 variables collected from 5 available tables.

- Coverage table. Data related to each policy on the coverage level: emission date, renovation date, changes over time.
- Object table. Variables related to the packages: entry date, changes over time, premiums, car-related variables.
- Simulation table. Data was collected during the simulation process offline and online. Contains car and customer-related variables.
- Policy characteristics. The table contains all actual data about each policy and each customer.
- Analytical Base table. Variables that describe customers and their relationships with the company. Include all historical data about each customer.

For this project, only the customers with specific characteristics were used.

Figure 3.1 shows the universe of the clients and the process of selecting the final observations.

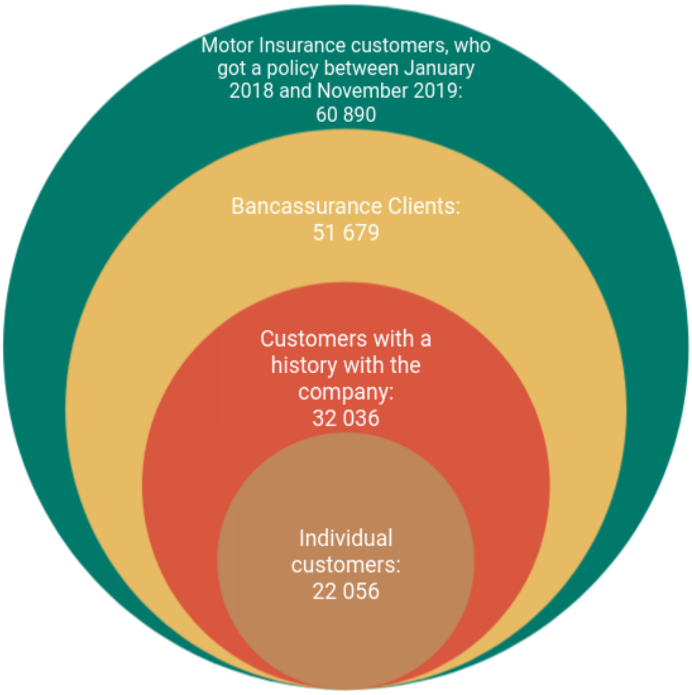

Motor Insurance customers, who got a policy between January 2018 and November 2019: 60 890
Bancassurance Clients: 51 679
Customers with a history with the company: 32 036
Individual customers: 22 056

Figure 3.1 - Data universe and selection of the final observations

In the end, the final dataset contained 22 056 observations before the data preparation step. All variables were aggregated into one table at the customer level.

### 3.3.2. Main Findings

After gathering the data, a simple data exploration analysis was performed. The figures below illustrate the main findings that were the most interesting for the business.

Among all the clients top 3 car brands are Renault, Peugeot, and Opel.

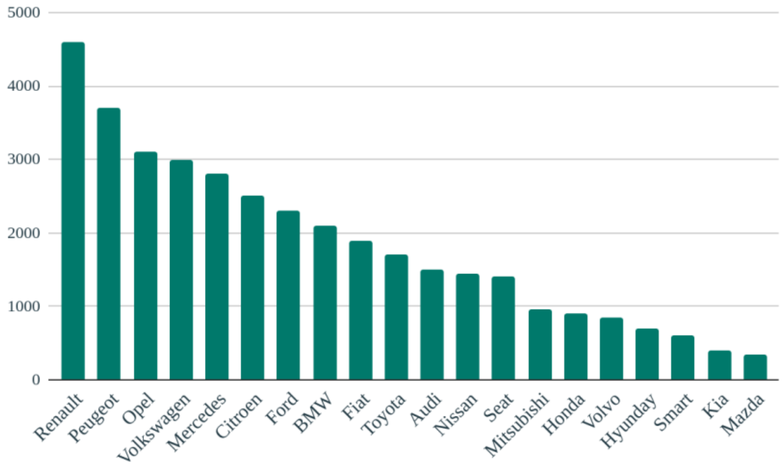


Figure 3.2 - Popularity of car brands

The majority of customers are new, but the distribution presented in Figure 3.3 proves that clients tend to stay with the company for years.

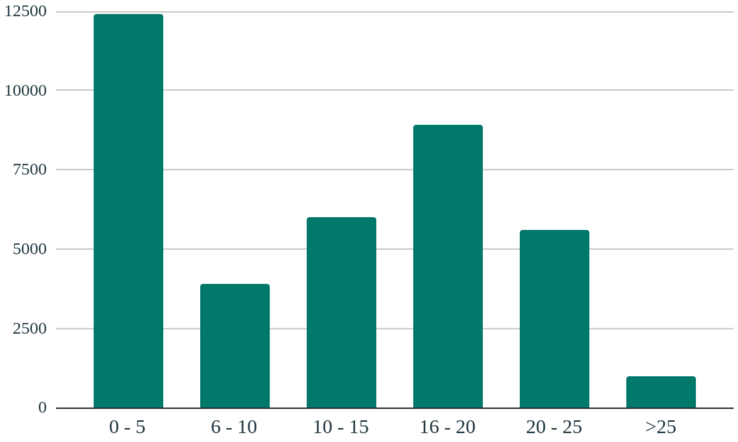


Figure 3.3 - Distribution of years as customers

The most exciting and important finding is the age of the clients. Young people tend not to buy this product. The project manager explained it with the higher price for the policy and mentioned that the company targets people older than 35.

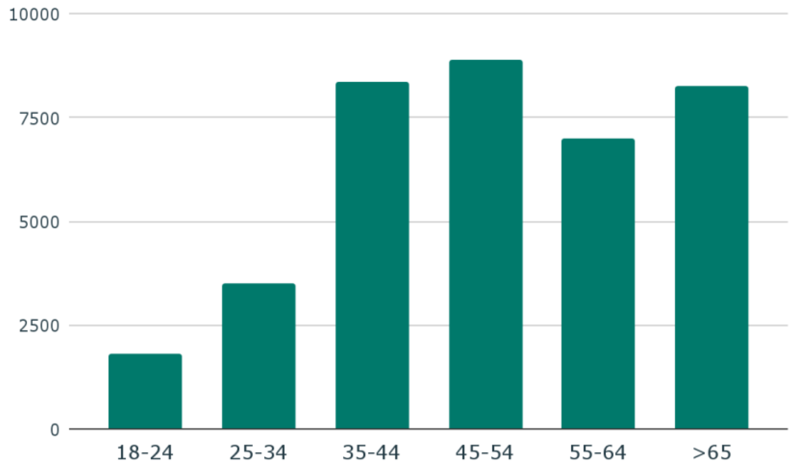


Figure 3.4 - Age distribution

### 3.3.3. Target definition

The target definition is the most critical step for this project because the target variable needed to be created based on existing variables.

It was decided to identify classes as combinations between packages and additional coverages that customers can have.

Customers can choose from three packages: Mini, Extra, Top, and extra thirteen coverages. Some of the coverages are available by default, some are optional, and some are unavailable in a specific package.

The main challenge of this step was to get from this complex approach with different combinations to a well-defined target.

The first step in defining classes was splitting each package into a tree that shows the distribution of policies between coverages.

In these trees, coverages are located depending on the number of policies that have them. Thus, the more popular the coverage, the closer it is to the root of the tree.

This approach was helpful to understand:

- Difference between customers who prefer each package;
- Number of policies per each package;
- Cut points for each package

At the end of this step, twenty-seven classes were defined.

Figure A2 in Appendix describes the graphic representation of this process.

The second step was to decrease the number of classes using the business approach.

Based on the product knowledge of the main stakeholders, Target 1 was defined, consisting of 11 classes split based on combinations of coverages for each package.

Table 3.2 shows the final classes defined after two steps of target creation.

| Package | Class | Coverages |
|---|---|---|
| **Extra** | Class 1 | ASB + QIV |
| | Class 2 | ASB + Mini |
| | Class 3 | AST + QIV |
| | Class 4 | AST |
| **Top** | Class 5 | CCC + VDS + VAV + (FDN or ADV) |
| | Class 6 | CCC + (VDS or VAV) |
| | Class 7 | rest |
| | Class 8 | CRO + VDS + VAV + (FDN or ADV) |
| | Class 9 | CRO + (VDS or VAV) |
| | Class 10 | rest |
| **Mini** | Class 11 | rest |

Table 3.2 - Target 1 Description

Since the beginning of the target definition step, it became clear that this project is not the same as most machine learning projects. Therefore, as well as testing different models and parameters and choosing the best one, it was necessary to define one target that would be equally good from both an analytical and business perspective.

The first defined target contained too many classes. As a result, the dataset was unbalanced, and most classes were not different. In order to achieve results that could be useful for business, it was needed to combine classes in different ways.

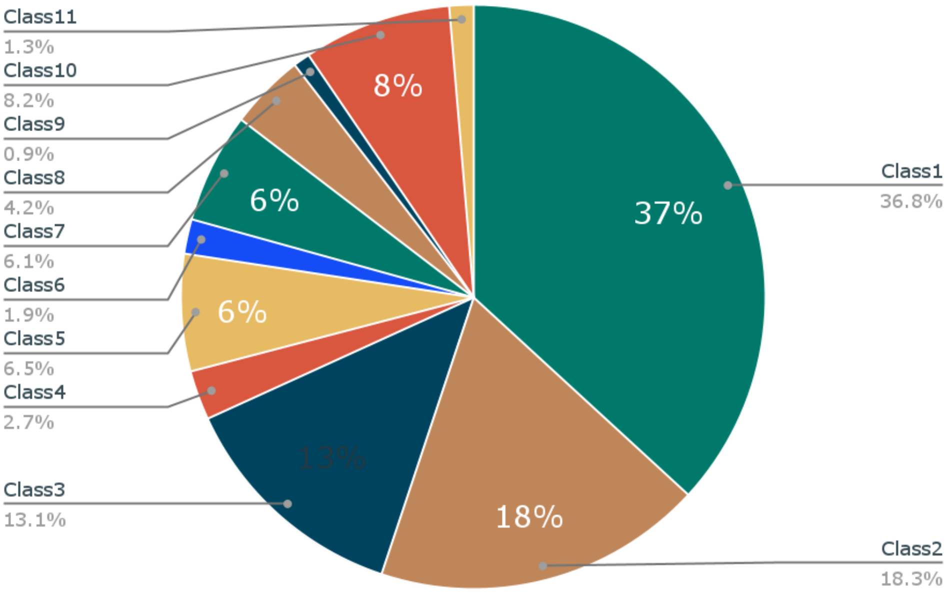


Figure 3.5 - Proportion of each class in the target variable

After a conversation with a product manager, 20 target variables were defined that contained different combinations of all classes.

A detailed table with all target variables can be found in Appendix A3.

## 3.4. DATA PREPARATION

In order to have a successful model, data preparation has to be performed. Data preparation is the process of manipulating data into a form that can be used for the following analysis.

This step was the most time-consuming of the whole project.

### 3.4.1. Missing values

Since the dataset was not big compared to other projects in the company, it was important not to lose any observations and to treat missing values.

There weren't a lot of variables with missing values. Based on the analysis and opinion of the stakeholders, for numeric variables, such as the age of a driver, the tree method was used. And for interval variables related to premiums paid, the mean method was used.

### 3.4.2. Outlier Handling

Outliers are values that are extremely large or extremely small. Some Machine Learning models are outlier sensitive; therefore, it was necessary to treat outliers in the dataset.

SAS Enterprise Miner has a feature to detect and handle outliers.

After the analysis and discussions with the stakeholders, Standard deviation from the Mean with a threshold equal to three was chosen as a limit method for outlier handling.

### 3.4.3. New variables

Sometimes during the data exploration and preparation, it gets clear that some variables cannot be used in the modeling process as they are. But, on the other hand, these variables can still be used to create new variables.

Based on the customers' historical data, four new variables related to motor insurance were created.

- Who had an active policy at the moment of purchase of a new one:
    a. Cannibalization: customers who canceled the policy one month before renewal and three months after purchasing the policy.
    b. Has a policy now: who has an active policy and issued a new one (no cancellation of old policy for at least more than three months after purchase)
- Who never had a policy before.
- Who canceled more than one month before issuing a new policy.

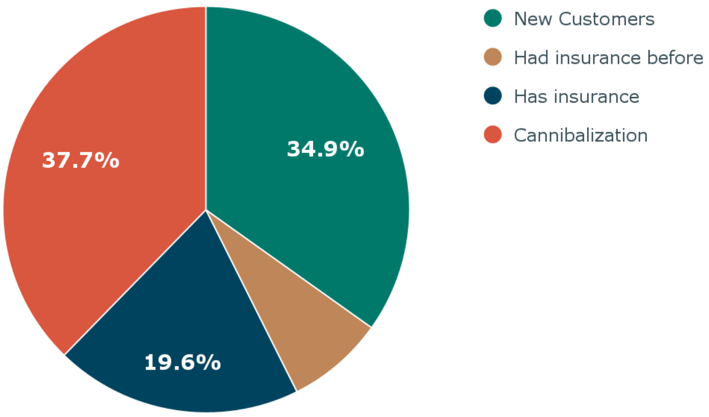


Figure 3.6 - Proportion of new variables in the dataset

Based on the historical data of the clients with other insurance in the company, three new variables were created:

- Customers who have one other type of insurance (LOB) with the company
- Customers who have more than one type of insurance (LOB) with the company
- Customers who don’t have any other type of insurance (LOB) with the company except motor insurance.

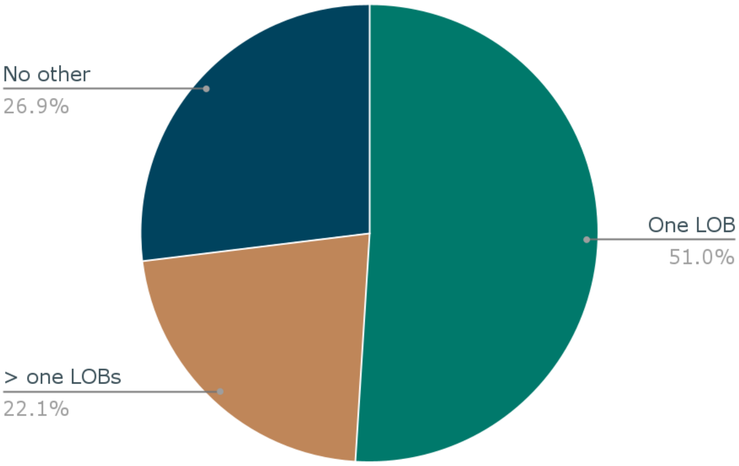


Figure 3.7 - Proportion of new variables in the dataset

### 3.4.4. Variable transformation

Transforming the variables before modeling can improve the predictive power of the models since transformations can cut off the noise. In this project, the Box-Cox transformation was used.

### 3.4.5. Variable selection

The variable selection step can improve the model's performance since it removes variables that are not relevant. In this project, the Decision Tree was chosen as a variable selection method.

The most important variables are presented below in the modeling part 3.5 table 3.4

## 3.5. Modeling

### 3.5.1. Dataset Preparation

During the modeling step, four algorithms suitable for multiclass classification were employed: Neural Network, Random Forest, Gradient Boosting, and Decision tree.

Before applying any of the models, the data went through some transformation.

- Partition. Data was split into test and validation datasets (80/20 ratio)
- Variable selection
- Variable transformation
- Decision weights for an imbalanced dataset

In the end, there were eight baselines with different parameters.

| | Baseline 1 | Baseline 2 | Baseline 3 | Baseline 4 | Baseline 5 | Baseline 6 | Baseline 7 | Baseline 8 |
|---|---|---|---|---|---|---|---|---|
| Variable Transformation | No | Yes | No | Yes | No | Yes | No | Yes |
| Variable Selection | Yes | No | No | Yes | Yes | No | No | Yes |
| Decision Weights | Yes | Yes | Yes | Yes | No | No | No | No |

Table 3.3 - Baselines Description

### 3.5.2. Parameters tuning

For each model, a parameter tuning process was applied.

#### 3.5.2.1. Decision Trees

Maximum Depth — specifies the maximum number of generations of nodes in the decision tree. The parameters tested: 6, 7, 8, 9.

To split interval variables, p-value was used.

To split nominal variables, Entropy was used.

#### 3.5.2.2. Random Forest

Number of variables specifies the number of input variables to consider splitting on in a node. The parameters tested: 6, 7, 8, 9, 10

Number of trees specifies the number of trees in the forest. Tested: 800, 900, 1000, 1200, 1500

#### 3.5.2.3. Gradient Boosting

Interaction depth specifies how many splits the model has to perform on a tree. Tested: 8, 9, 10, 12

Number of trees in the model: 700, 1000, 1200, 1500

Shrinkage is used to specify how much to reduce the prediction of each tree. Tested: 0.01, 0.03, 0.05, 0.07, 0.09, 0.11

Minimum observations in a node in the model: 10, 20, 30

#### 3.5.2.4. Neural Network

Number of hidden layers: 0, 1, 2, 3.

### 3.5.3. Results

Due to the imbalance of the dataset, the Weighted F1 Score was selected to measure the performance of each model.

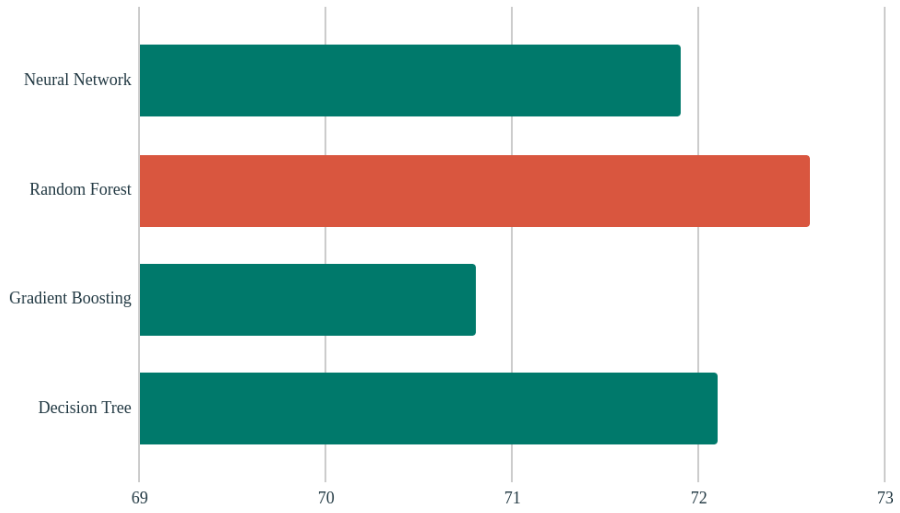


Figure 3.8 - F1 Score, % of the best combinations of parameters for each model

As seen in Figure 3.8, there was not much variability between the models; they showed almost the same results.

In the end, the model with the best performance was Random Forest with Trees = 800, Variables = 8, and Baseline 5 with the F1 Score = 72.6

Results for each model and baseline can be found in Appendix B2.

The Random Forest model is not needed to use variable selection since the algorithm chooses them itself. The most important variables are presented in Table 3.4.

| Variables describing a car | Variables describing customers |
|---|---|
| Car Age | Other LOBs |
| Car brand | Driver Age |
| Vehicle Potential | Segment Id |
| Vehicle Cylinder | Unsuccess Sales Campaign |
| Number of doors | Segment |
| | Number of active car policies |
| | Number of LOBs ever |

Table 3.4 - Variables used for the final Model

There are 13 variables. The proportion of variables describing cars and variables related to customers are 46% to 54%.

### 3.5.4. Target selection

As mentioned above, one of the biggest challenges during this project was to define the target and split the dataset into classes that would be meaningful from both analytical and business perspectives.

Figure 3.9 shows F1 Scores for most defined and tested targets and the number of classes in each target. So naturally, the more classes were in the target, the wider variety of clients were there. But in the end, it was necessary to make a trade-off between analytics and business.

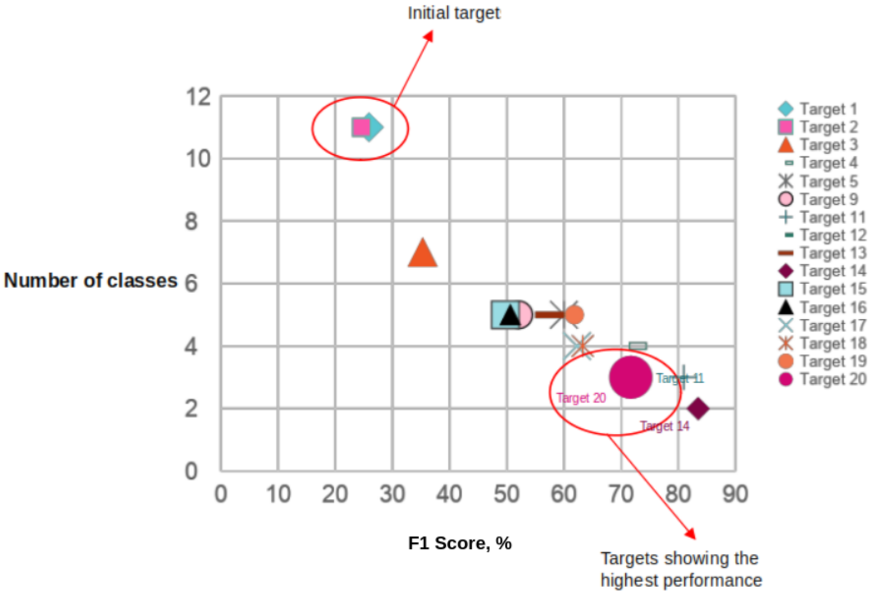


Figure 3.9 - Graphical representation of the number of classes and F1 Score for each Target Variable

The best performance was shown by Target 11, Target 14, and Target 20. The difference between them was the splitting method and the number of classes.

As shown in Figure 3.9, the target with the best performance was Target 14, with the best F1 Score equal to 84.5%. Even though it made sense from the analytical point of view to choose this target variable, from the business point of view, this target variable doesn't have much meaning since it only contains two classes, and there is no distinction between packages.

| Target 14 | | |
|---|---|---|
| Package | Class | Coverages |
| Extra + Mini | Class1 | All |
| Top | Class2 | All |

Table 3.5 - Target 14 Description

After meeting with the stakeholders, Target 20 was chosen as the most reasonable for this project from analytical and business perspectives.

| Target 20 | | |
|---|---|---|
| Package | Class | Coverages |
| Extra | Class 1 | (ASB or AST) + QIV |
| | Class 2 | (ASB or AST) & Mini + ODV + Class11 |
| Top | Class 3 | All |

Table 3.6 - Final Target Description

## 3.6. EVALUATION

In order to validate the model, backtesting has been performed. It was applied to historical data from December 2019 until the middle of May 2020.

The backtesting on the historical data had the F1 Score = 72.3, which is only 0.3, smaller than Test F1 Score.

The dataset for backtesting includes customers who opened a motor policy during a period of COVID-19 quarantine. People were forced to stay at home; therefore, it was assumed that customers' behavior changed.

In the end, it was decided to perform backtesting on three datasets:

- Data before COVID-19, from December 2019 until February 2020
- Data during Covid-19, from March 2020 until the middle of May 2020
- Total dataset including observations from December 2019 until middle May 2020.

As a result, it showed that the first test group performed a little better than the two others, although the results of all three datasets were very similar.

Figure 3.10 presents the final evaluation results of the model.

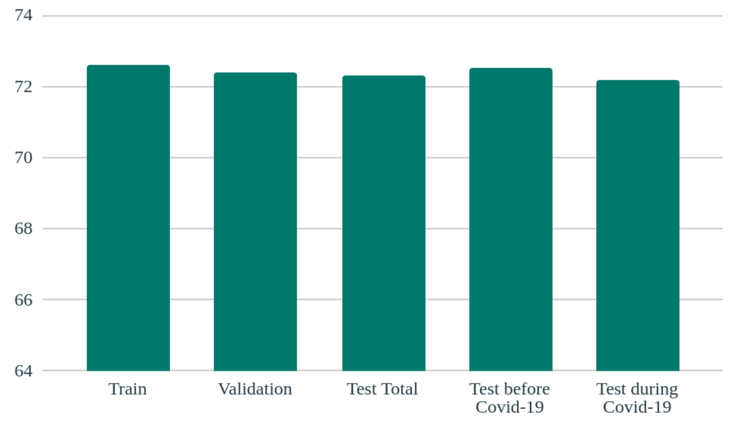


Figure 3.10 - F1 Score of train, validation and test datasets, %

## 3.7. Deployment

Deployment is the next logical step that is following the evaluation of the model.

This project has four ways of implementing it in the company.

- Previous simulations: to score the data available from previous simulations.
- Winback: using the model on the data of customers who once had a policy with the company but canceled.
- Simulation process: Use model insights to define rules for package recommendation during simulation in real-time.
- Digital: Implementing the model during simulation in digital products.

During the period of the project, the final model was prepared for deployment. However, the next step, exploring the possibilities of deployment with stakeholders, was not started due to COVID-19 complications.

## 4. CONCLUSIONS

This report describes the process of solving a multiclass classification problem.

The project started with a business understanding where the main objectives were identified.

The next step was data understanding. The data was collected from different data sources, and the target variable was identified. Based on the business and data analysis, the initial target was regrouped, and in the end, there were 20 target variables identified. Each target variable contains a different number of classes based on various combinations of packages and coverages of motor insurance. Therefore, it was essential to understand which target has the best performance and makes the most sense for business.

The following step was data preparation, during which data was cleaned, and new variables were manually created. Before starting the modeling part, it was also important to make variable transformation and variable selection to get better results during the next part.

The main part of this project was modeling. Four models were tested to solve the multiclass classification problem. F1 Score was chosen as a performance measure. The model with the biggest F1 score is Random Forest. The last task to do was to select the target that was a compromise between business and analytics. In the end, it was decided to use the target with three classes.

The next step after modeling is always evaluation. During this step, it was proven that the chosen model shows the same results on the test dataset, despite the influence of the Covid-19 lockdown that happened during the period of this project.

Finally, four methods of deployment were developed, and the model was prepared for future implementation.

# 5. LIMITATIONS, RECOMMENDATIONS FOR FUTURE WORK, AND MAIN LEARNING

## 5.1. LIMITATIONS AND RECOMMENDATIONS

During the project, some limitations were identified and listed below, along with recommendations for future work:

- Not all relevant variables were available to use during the project for different reasons. Therefore, the recommendation here is to use other variables that can be relevant for the model. These variables can be created using information from other sources (for example, customer-related variables: family units, other insurance products) or obtained from the partner bank in anonymized form.
- The data set was imbalanced, so other balancing methods can be applied—for example, resampling. Instead of creating one perfect target variable and making the number of classes smaller, it would be interesting to use other balancing methods and make a comparison.
- In order to gather as much data as possible, information about offline and online simulations was used together. In the future, it would be interesting to compare these two in order to understand if online behavior is different from offline.

## 5.2. MAIN LEARNINGS

After working on the project, a few main learnings were made:

- Planning. It is crucial to have more time for each step of the project if something goes wrong. Split one big task into small ones and define deadlines for them.
- Finding a compromise between data science & business. Of course, academic knowledge is important, but there should be a balance between business and technology.
- Asking for help. Don’t be afraid to ask for help, don’t try to deal with everything if the project gets stuck or there are doubts.

# 6. BIBLIOGRAPHY

Advanced Business Analytics (2012). SAS Institute, Inc.

Berry, M., Linoff, G. (2009). Data Mining Techniques: Theory and Practice, 19-141.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A. (1984). Classification and Regression Trees. *Boca Raton, FL: CRC press.*

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (1999, 2000). CRISP-DM 1.0.

Conway, D., Myles White, J. (2012). Machine Learning for Hackers: Case Studies and Algorithms to Get You Started, *1st Edition.*i

Dorogush, A. V., Ershov, V., Gulin, A. (2018). CatBoost: gradient boosting with categorical features support

Fernandez, A., Garcia, S., Galar, M., Prati, R., Krawczyk, B., Herrera, F. (2018). Learning from Imbalanced Data Sets, *Springer,* 197-221.

Friedman, J. (1999). Greedy Function Approximation: A Gradient Boosting Machine.

Goutte, C., Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation, *Prentice Hall, New Jersey*

He, H., Ma, Y., (2013). Imbalanced Learning: Foundations, Algorithms, and Applications, *Wiley.*

Kamakura, W. (2007). Cross-selling: Offering the right product to the right customer at the right time.

Kearns, M. (1988). Thoughts on Hypothesis Boosting.

Kolo, B. (2010). Binary and Multiclass Classification

Kuhn, M., Johnson, K. (2018). Applied Predictive Modeling. *Springer,* 198-205, 419-442.

Michie, D., Spiegelhalter, D. J.m, Taylor, C.C. (1994). Machine Learning, Neural and Statistical

Classification

Mitchell, T., Hill, M. (1997). Machine Learning.

Molnar, C. (2020). Interpretable Machine Learning on Decision Tree, Chapter 4.

Molnar, C. (2020). Interpretable Machine Learning on Decision Tree, Chapter 4.

Nabi, J. (2018). Machine Learning — Multiclass Classification with Imbalanced Dataset. Retrieved from towardsdatascience.com.

Nagpal, A. (2017). Decision Tree Ensembles- Bagging and Boosting. Retrieved from towardsdatascience.com.

Natekin, A., Knoll, A. (2013). Gradient boosting machines, a tutorial. Retrieved from www.frontiersin.org

Nisbet, R., Elder, J., Miner, G. (2009). Handbook of Statistical Analysis and Data Mining Applications, Chapter 10.

Novaković, J. Dj., Veljović, A., Ilić, S. S., Papić, Z., Tomovic, M. (2017). Evaluation of Classification Models in Machine Learning.

Ramageri, B. M. (2010). Data Mining Techniques and Applications.

Rodne, L. (2009). Introduction to Insurance.

Rygielski, C., Wang J. C., Yena, D. C. (2002). Data mining techniques for customer relationship management.

Safavian, R., S., Landgrebe, D., (1991) A Survey of Decision Tree Classifier Methodology.

SAS Enterprise Guide. Fact Sheet (2017).

SAS Enterprise Miner. Fact Sheet (2017).

SAS Enterprise Miner Reference Help.

SAS Enterprise Miner Version 14.3, (2020).

SAS Help Center: Introduction to SEMMA.

Shmueli, B. (2019). Multi-Class Metrics Made Simple, Part II: the F1-score. Retrieved from towardsdatascience.com.

Silipo, R. (2020). Ensemble Models: Bagging & Boosting. Retrieved from medium.com.

Silva-Palacios, D.,Ferri, C., Ramírez-Quintana, M. J. (2017). Improving Performance of Multiclass Classification by Inducing Class Hierarchies

Terry-Jack, M. (2019). Tips and Tricks for Multi-Class Classification. Retrieved from medium.com.

Witten, I. (2020). Data Mining: Practical Machine Learning Tools and Techniques, *4th Edition.*

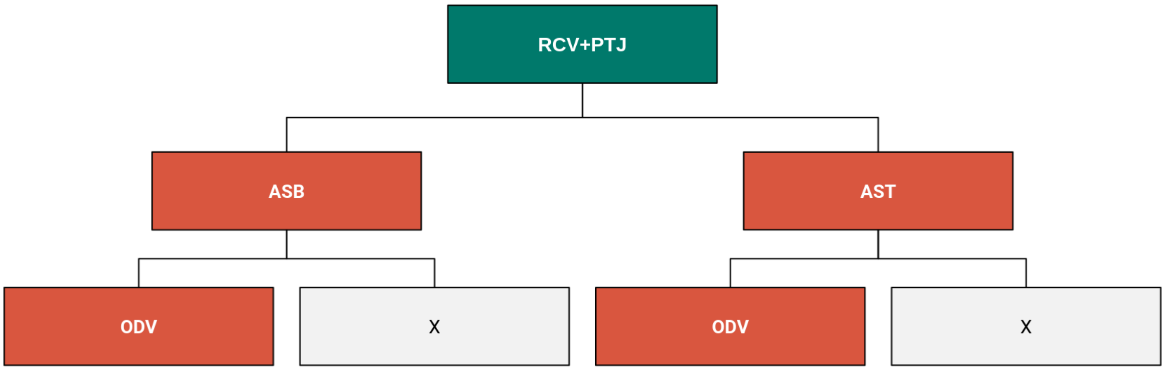
# 7. APPENDIX A. DATASET

## A1. Packages and Coverages

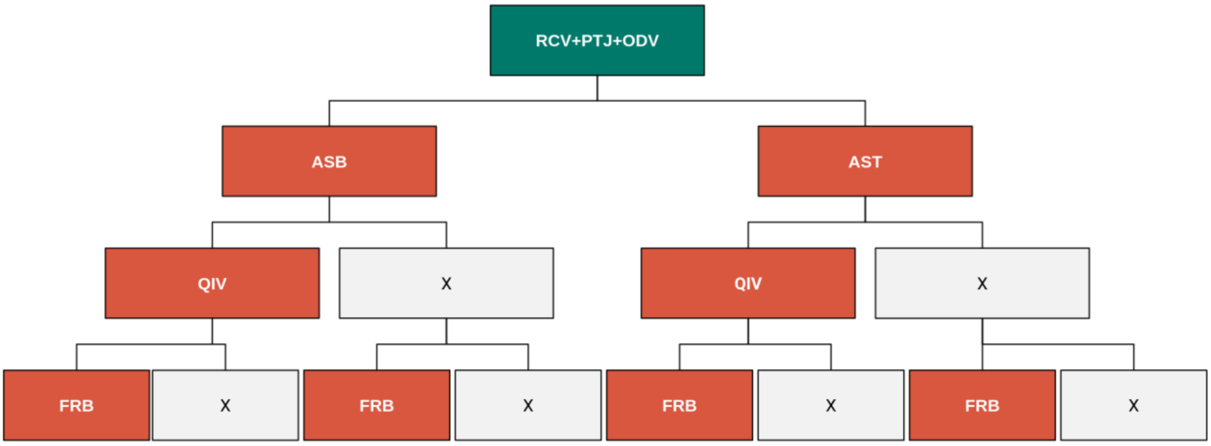
| Coverage | Abbr | Mini | Extra | Top |
|---|---|---|---|---|
| Third party liability | RCV | ✓ | ✓ | ✓ |
| Travel Assistance | ASB/AST | Base | Base or Total | Base or Total |
| Legal Protection | PTJ | ✓ | ✓ | ✓ |
| Vehicle Occupants | ODV | Optional | ✓ | ✓ |
| Isolated Breakage of Glass | QIV | X | Optional | ✓ |
| Crash, Collision, Roll-Over | CCC/CRO | X | X | ✓ |
| Fire, Lightning Strike, Explosion | IRE | X | Optional | ✓ |
| Theft or Robbery | FRB | X | Optional | ✓ |
| Natural Phenomena | FDN | X | Optional | Optional |
| Acts of Vandalism | ADV | X | X | Optional |
| Replacement Vehicle - Accident | VDS | X | Optional | Optional |
| Replacement Vehicle - Breakdown | VAV | X | Optional | Optional |
| Baggage | BAG | X | X | Optional |
| Pack Relax | PMU | Optional | Optional | Optional |
| Pack Vintage | PSR | Optional | Optional | Optional |
| Pack Prestige | PPE | Optional | Optional | Optional |

## A2. Defining classes

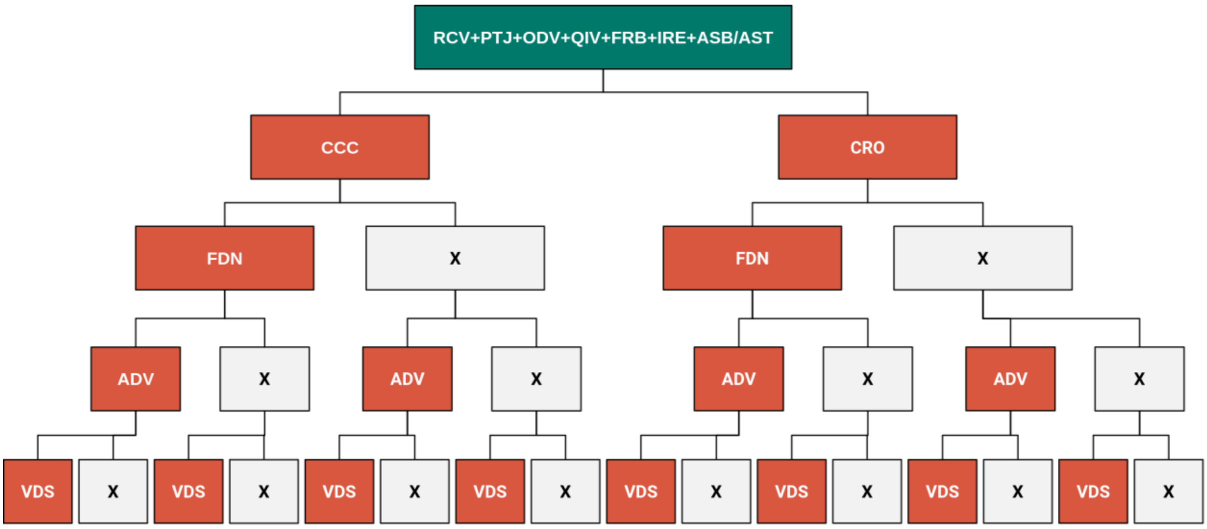
### A2.1. Package “Mini”



### A2.2. Package “Extra”



### A2.3. Package “Top”



## A3. Target variables

| Class | Package | Target1 Classes |
|---|---|---|
| Target **3** | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2 + Class4 |
| Class3 | Top | Class5 + Class6 |
| Class4 | | Class 7 |
| Class5 | | Class 8 + Class9 |
| Class6 | | Class10 |
| Class11 | Mini | Class11 |
| Target 4 | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2 + Class4 |
| Class3 | Top | All Classes |
| Class11 | Mini | Class11 |
| Target 5 | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2 + Class4 |
| Class3 | Top | Class5 + Class7 |
| Class4 | | Class 6 + Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 6 | | |
| Class1 | Extra | Class1 + Class2 |
| Class2 | | Class3 + Class4 |
| Class3 | Top | Class5 + Class7 |
| Class4 | | Class 6 + Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 7 | | |
| Class1 | Extra | Class1 + Class2 + Class11 |
| Class2 | | Class3 + Class4 |
| Class3 | Top | Class5 + Class7 |
| | | Class 6 + Class 8 + Class 9 + Class 10 |
| | Mini | Class11 |
| Target 8 | | |
| Class1 | Extra | Class1 + Class2  Class4 |
| Class3 | | Class3 |
| Class2 | Top | Class5 + Class7 |
| Class4 | | Class 6 + Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 9 | | |
| Class1 | Extra | Class1 + Class2 |
| Class2 | | Class3+Class4 |
| Class3 | Top | Class5 + Class6 + Class7 |
| Class4 | | Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 10 | | |

| | | |
|---|---|---|
| Class1 | Extra | Class1 + Class2+Class4 |
| Class2 | | Class3 |
| Class3 | Top | Class5 + Class6 + Class7 |
| Class4 | | Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 11 | | |
| Class1 | Extra | All |
| Class2 | Top | All |
| Class11 | Mini | All |
| Target 12 | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2+Class4 |
| Class3 | Top | Class5 + Class6 + Class7 |
| Class4 | | Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 13 | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2+Class4 |
| Class3 | Top | Class5 + Class6 + Class8 + Class9 |
| Class4 | | Class7 + Class 10 |
| Class11 | Mini | Class11 |
| Target 14 | | |
| Class1 | Extra + Mini | All |
| Class2 | Top | All |
| Target 15 | | |
| Class1 | Extra | Class1 + Class2 + Class4 |
| Class2 | | Class3 |
| Class3 | Top | Class5 + Class6 + Class7 |
| Class4 | | Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 16 | | |
| Class1 | Extra | Class1 + Class2 |
| Class2 | | Class3 + Class4 |
| Class3 | Top | Class5 + Class6 + Class8 + Class9 |
| Class4 | | Class7 + Class 10 |
| Class11 | Mini | Class11 |
| Target 17 | | |
| Class1 | Extra | All |
| Class2 | Top | Class5 + Class6 + Class8 + Class9 |
| Class3 | | Class7 + Class 10 |
| Class11 | Mini | Class11 |
| Target 18 | | |
| Class1 | Extra | All |
| Class2 | Top | Class5 + Class6 + Class7 |
| Class3 | | Class 8 + Class 9 + Class 10 |
| Class11 | Mini | Class11 |
| Target 19 | | |
| Class1 | Extra | Class1 + Class2 |
| Class2 | | Class3 + Class4 |
| Class3 | Top | All |
| Class11 | Mini | Class11 |
| Target 20 | | |
| Class1 | Extra | Class1 + Class3 |
| Class2 | | Class2 + Class4 + Mini |
| Class3 | Top | All Classes |

# 8. APPENDIX B. MODELING

B1. Data Baselines

| | Baseline 1 | Baseline 2 | Baseline 3 | Baseline 4 | Baseline 5 | Baseline 6 | Baseline 7 | Baseline 8 |
|---|---|---|---|---|---|---|---|---|
| Variable Transformation | No | Yes | No | Yes | No | Yes | No | Yes |
| Variable Selection | Yes | No | No | Yes | Yes | No | No | Yes |
| Decision Weights | Yes | Yes | Yes | Yes | No | No | No | No |

B2. Modeling Results

| | Baseline 1 | Baseline 2 | Baseline 3 | Baseline 4 | Baseline 5 | Baseline 6 | Baseline 7 | Baseline 8 |
|---|---|---|---|---|---|---|---|---|
| Decision Tree | 71.29 | 71.24 | 71.29 | 71.24 | 72.12 | 71.19 | 72.12 | 71.19 |
| Gradient Boosting | 69.81 | 69.47 | 69.81 | 69.47 | 70.08 | 70.79 | 70.08 | 70.79 |
| Random Forest | 70.25 | 68.95 | 71.19 | 69.48 | 72.60 | 71.08 | 71.92 | 71.82 |
| Neural Network | 69.12 | 69.50 | 69.57 | 69.78 | 71.96 | 70.56 | 71.96 | 70.98 |