Data Mining Report

Assignment Resit, June 2023

University of Stirling

CSCU9M5

# Business Understanding

The task at hand consists in building a classifier for a car insurance company that predicts if a client will claim insurance or not, with the end goal being to develop an ML model that uses customer details as input and outputs binary predictions of whether customers will make insurance claims or not.

The data provided in the *resit\_dataset\_prepared.csv* file consists of 1000 records of customer information from the car insurance company. The data set includes the following relevant attributes: gender, age, race, driving experience, education, income, credit score, vehicle details, marital status, number of children, country, post code, annual mileage, vehicle type, speeding violations, driving under influence, and past accidents. The target variable is *HAS\_CLAIMED\_INSURANCE* which indicates if a customer has claimed insurance or not.

The finished system will require a classifier model capable of predicting whether or not a customer will claim insurance based on the attributes listed above. The system will take into consideration customer details (input) and will produce a binary prediction (output) of “claim” (1) or “no claim” (0).

This task suits the requirements of a Data Mining Approach due to: its large dataset which provides a substantial amount of data to extract patterns and insights; its complex relationships in insurance claiming decisions which tend to depend of factors such as customer demographics, driving history, and policy details, which are analysed through data mining techniques, thus finding patterns that are hard to find and discern manually; its classification problem since predicting whether or not a customer will claim insure is a classification problem, and a common application of data mining techniques; and its decision support to develop an accurate classifier which will help the car insurance company to assess risks, optimize premiums, and make informed decisions based on customer data.

The following terminology can be found throughout the report:

* Attribute: also known as feature or variable, it is a specific characteristic or property of data. Within this report, attribute will refer to the various attributes of the csv file dataset which help describe and categorize the customers.
* Binary prediction: prediction or classification with two and only two outcomes or categories. Within this report, binary prediction will refer to the model’s prediction as to whether or not the customer will claim insurance. It will be represented as “1” for “claim” and “0” for “not claim”.
* Input: within this report, input will refer to the customer attributes provided in the csv file to be used in the model for prediction.
* ML Model (or simply Model): a mathematical representation of the relationships between input attributes and the target variable.
* Output: prediction or result generated by a ML model. Within this report, output will refer to the binary prediction of whether or not a customer will claim insurance.
* Target variable: also known as dependent variable or response variable, it is the variable that will be predicted. Within this report, target variable will refer to the “HAS\_CLAIMED\_INSURANCE” column within the csv file which indicates whether or not a customer has claimed insurance.
* Task: the task at hand, to build a classifier ML model that predicts whether or not a customer will claim insurance based on their attributes as listed in a csv file.
* Variable: an attribute or feature of a dataset. E.g.: age, gender, country, …

With regards to project methodology, I will be using the Cross-Industry Standard Process for Data Mining (CRISP-DM) which will ensure a structured and systematic approach to the project, allowing me to conduct reliable analysis and to make informed decision-making based on the output of the predictive model. CRIPS-DM consists of the following 6 stages:

1. **Business Understanding:** acknowledging business goals and defining the problem.
2. **Data Understanding:** exploring and understanding the data provided. Assess data quality and identify relevant attributes.
3. **Data Preparation:** readying data for modelling by addressing missing values, data cleaning, normalisation, feature selection, and data splitting into the training, validating, and testing datasets.
4. **Modelling:** choosing the appropriate ML model algorithms, training them using the training dataset, and tuning hyperparameters for optimal performance.
5. **Evaluation:** evaluating the model’s performance through appropriate evaluation metrics and selecting the best-performing model.
6. **Deployment:** documenting findings, preparing a summary of the project, and presenting said report to stakeholders.