Capstone Project: Predicting and Explaining Caravan Policy Ownership

Manohari Wijesooriya

Student ID: 501212269

The Chang School of Continuing Education, Toronto Metropolitan University

CIND820 Big Data Analytics Project

Dr. Tamer Abdou

May 15, 2023

Contents

1. Abstract

The aim of this project is to predict if a customer will purchase a Caravan Insurance Policy based on socio-demographic and product ownership data in an insurance company.

Cross-selling involves selling complementary products to existing customers. This is a business case to find a machine learning solution to support cross selling of insurance product. It is about predicting who would be interested in buying a caravan insurance policy and to give a relevant explanation. If the company had a better understanding of who their potential customers were, they would know more accurately who to send policy quotes to, so some of this waste and expense could be reduced.

The main business problem:

- Can you predict who would be interested in buying a caravan insurance policy and give an explanation why?

The business problem is broken down to following research questions

- Predict which customers are potentially interested in a caravan insurance policy.
- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

Additional Research questions

- How Caravan Insurance ownership does varies across different demographic areas, and can we create distinct profiles of Caravan Insurance customers based on sociodemographic data?
- predicting a customer's likelihood to purchase Caravan Insurance based on their sociodemographic characteristics
- What frequent associations can be identified in the product ownership data?

Dataset: **Insurance Company Benchmark (COIL 2000).** This data set used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data consists of 86 variables and includes product usage data and socio-demographic data.

Dataset can be found in this link:

https://archive.ics.uci.edu/dataset/125/insurance+company+benchmark+coil+2000

TICDATA2000.txt: training dataset with 5822 records and 86 attributes

TICEVAL2000.txt: test dataset without the target; includes 4000 records and 85 attributes

TICTGTS2000.txt – targets for test dataset; includes 4000 records and 1 attribute

Each record of the file consists of 86 attributes, containing sociodemographic data (attribute 1-43) and product ownership (attributes 44-86).The sociodemographic data is derived from zip codes. All customers living in areas with the same zip code have the same sociodemographic attributes. Attribute 86, "CARAVAN: Number of mobile home policies", is the target variable.

In this research classification analysis will be used for the prediction part. First, classification algorithms Multinomial Naïve Bayes, Decision Tree, Random forest, Logistic regression, K Nearest Neighbours will be used in modelling. Then results of these models will be compared using evaluation matrices accuracy, precision and recall. The best model will be the model with high recall on target = 1 and the high accuracy.

The purpose of the description task is to give a clear insight to why customers have a caravan insurance policy and how these customers are different from other customers. Descriptions can be derived using Apriori and Fpgrowth algorithms. In this task rules will be set to identify Caravan insurance customer and measure statistical significance of those rules.

This analysis will be mainly done using Python and use Google Collab to execute the code.

Other tools R, SAS and Tableau will be used in exploratory data analysis and prepare visualizations.