# Capstone Project: Predicting and Explaining Caravan Policy Ownership

## Literature Review

Manohari Wijesooriya

Student ID: 501212269

The Chang School of Continuing Education, Toronto Metropolitan University

CIND820 Big Data Analytics Project

Dr. Tamer Abdou

June 9, 2023

## Contents

# 1. Introduction

A Norwegian insurance company was interested in a machine learning solution to find best customers to market its caravan insurance product. Without sending mass email to all customers, it is cost effective to identify best possible customers who will buy a caravan insurance and market only for those.

## 1.1. Statement of the Problem

The main research question:

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

Following tasks will be performed in this analysis.

- 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?

## 1.2. Background

The data was supplied by Sentient Machine Research. https://www.smr.nl/

This dataset is offered in a competition 'CoIL Challenge 2000'. The CoIL Challenge was a datamining competition organized by the Computational Intelligence and Learning Cluster, a network of excellence sponsored by the EU. It was held in the period of March-May 2000, in total 43 solutions were submitted.

## 2. Literature Review

Several articles were reviewed to gather efforts made by previous researches analysing this dataset.

Charles et al., 2000, the first price winner of the completion in the prediction task used Naïve Bayes algorithm and identified 121 caravan policy holders out of 238 actual counts [2]. He has identified the strongest single predictor of having a caravan insurance policy is having a single car insurance policy where the contribution is high (level 6), or having two car policies[1] He has derived some attributes and used Boosting model.

I am planning to use Naïve Bayes and improve the information gain by combining attributes.

I read through the article from YoungSeong et al., 2000, the winners of the description task of the modelling competition [3]. They have used a combine method of artificial neural networks (ANNs) for prediction with evolutionary search for choosing the predictive features. The feature subset uses Evolutionary Local Search Algorithm (ELSA).They have considered distribution of each feature, normalized to the size of smaller one and a Chi-square test performed to see if the distributions were significantly different. They also conducted a search for simple association rules that would predict the purchase of a caravan policy. They have concluded contribution to the car policy is the strongest predictor.

In my research, I will use traditional machine learning algorithms such as Multinomial Naïve Bayes, Decision Tree, Random forest, Logistic regression, K Nearest Neighbours.

In his article Alexander et al. 2000, explains use of Python weka package in predicting caravan customers. He explains after removing duplicates and removing low information attributes, he could increase the accuracy of the model [4].

I will be using Naïve Base algorithm in Python in predicting caravan customers.

I reviewed article by Karishma et al. (no date). In her research, she managed to predict 130/238 customers correctly [5] using Naïve Bayes with bagging.

I will follow her data manipulation technique of re-coding categorical values with the mid value of the original range of value.

## References

[1] Charles Elkan. (2000). COIL CHALLENGE 2000 ENTRY. 1 - 2

http://www.liacs.nl/~putten/library/cc2000/ELKANP~1.pdf. Retrieved on May 25, 2023

[2] Charles Elkan. (2013). Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000. 1 – 5 Article 10.1145/502512.502576

https://www.researchgate.net/publication/2368301_Magical_Thinking_in_Data_Mining_Lessons_From_CoIL_Challenge_2000. Last accessed on July 18, 2023

[3] YoungSeong Kim and W.N. Street.(2000). CoIL Challenge 2000: Choosing and Explaining Likely Caravan Insurance Customers.

http://www.liacs.nl/~putten/library/cc2000/STREET~1.pdf. Last accessed on July 18, 2023

[4] Alexander K. Seewald. (2000). CoIL Challenge 2000 Submitted Solution.

http://www.liacs.nl/~putten/library/cc2000/SEEWAL~1.pdf. Last accessed on July 18, 2023

[5] Karishma Dudani. (no date) . Predicting Sale of Caravan Insurance Policy. 3 – Data Manipulation.

https://beginanalyticsblog.wordpress.com/2017/03/25/predicting-sale-of-caravan-insurance-policy/. Last accessed on July 10, 2023

## 3. Data and Information

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

The dataset consists of 86 attributes and 9822 data points. It is further divided into a training set (5822 observations) and a test set (4000 observations). Out of 86 attributes 2 are categorical (customer sub type, customer main type), 84 are numerical.

Refer to the data dictionary in the appendix. The dataset 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.

## 4. Exploratory Data Analysis

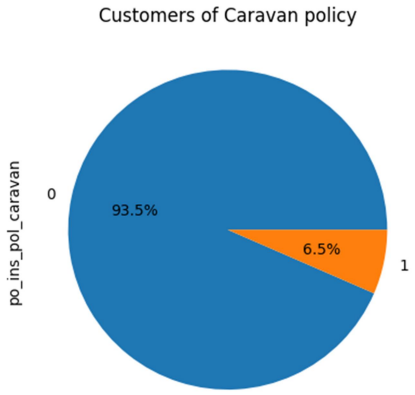
Complete exploratory data analysis can be found in this Github link.

https://github.com/manohariw44/TMU-Big-Data-Analytics-Capstone-Project/blob/d8539adaffffa6296d62935007c45ca18813fd4d/Pandas%20Profiling%20Report%20%E2%80%94%20Variable%20profile2%20-%20no%20duplicates%20.html

Link to correlation matrix

https://github.com/manohariw44/TMU-Big-Data-Analytics-Capstone-Project/blob/7bab4943c5656927af865cb47f91ddb3787deccb/corr_matrix_all%20attributes.png

There is small proportion (6.5%) of success targets in the validation dataset.




| target | # records |
|---|---|
| 0 | 4880 |
| 1 | 340 |
| | 5220 |

6.5%

## 5. Approach

Initially, I will spend some time in cleaning the dataset. I will prepare exploratory data analysis report, examine attributes and clean further by dropping attributes providing low information gain. I will combine some attributes meaningfully and derive new attributes.

The training dataset has small number of targets (6.5%) equal to 1. I will use over sampling technique; Synthetic Minority Oversampling Technique (SMOTE) by duplicating the rare class of the training dataset. I will also use k-fold cross validation method to avoid over or under fitting.

I will use Naïve Bayes algorithm for prediction task. I will also use algorithms Decision Tree, Random forest, Logistic regression, K Nearest Neighbours. Then results of these

models will be compared (accuracy, precision and recall will be examined) and select the best algorithm to predict caravan insurance customer.

For the description task, I will use Apriori, Fpgrowth and k-means algorithms.

**Step1 - Data Collection**
Upload datasets to Google Collab.

**Step 2 - Data Preparation**

- ∞ Wrangle data and prepare it for training
- ∞ Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, and data type conversions, etc.)
- ∞ Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- ∞ Split into training and evaluation sets

**Step3 - Choose a Model**

I will few algorithms to train model:Naïve Bayes, sklearn, Decision Tree, Random Forest, Balanced Random Forest, logistic regression and k neighbors.

**Step4 - Evaluate the Model**

I will run the model on evaluation dataset and evaluate model using matrices accuracy, precision and recall. The best model will be the model with high recall on target = 1 and the high accuracy.

**Step5 - Parameter Tuning**

I will use backward elimination method in selecting essential features of data.

I will do steps 3 – 5 iteratively over different algorithms, using different features and note down accuracy, precision and recall from each model. Then I will choose the best performing algorithm.

I am planning to use programming languages and tools in this research as follows.

| Task | Tool/package/ library |
|---|---|
| Data profiling, visualisation, feature engineering | Python packages : pandas, numpy, matplotlib, seaborne, pandas profiling |
| Data visualization | Tableau<br>R libraries : ggplot2 |
| Classification model | Python packages: sklearn<br>R libraries: ISLR, class |
| Explain predictions | apyori, Fpgrowth |