

Predicting and Explaining Caravan Policy Ownership

CIND820: Capstone Project

Project by: Manohari Wijesooriya

swijesooriya@torontomu.ca

Student #501212269

Supervisor: Dr. Tamer Abdou

Date of Submission: July 20, 2023

Table of Contents

[Abstract 3](#_Toc140786748)

[Introduction 4](#_Toc140786749)

[Context and Objective 4](#_Toc140786750)

[Approach 4](#_Toc140786751)

[Literature Review 5](#_Toc140786752)

[Background 5](#_Toc140786753)

[Recent Studies 5](#_Toc140786754)

[Data Collection and Analysis 6](#_Toc140786755)

[Data Source 6](#_Toc140786756)

[Exploratory Data Analysis 6](#_Toc140786757)

[Data Preprocessing and Preparation 7](#_Toc140786758)

[Data Cleaning 7](#_Toc140786759)

[Feature Engineering / Selection 7](#_Toc140786760)

[Methodology 8](#_Toc140786761)

[Predictive modelling approach 8](#_Toc140786762)

[Multinomial Naïve Bayes 8](#_Toc140786763)

[Decision Tree 8](#_Toc140786764)

[Random forest 8](#_Toc140786765)

[Logistic regression 8](#_Toc140786766)

[K Nearest Neighbours 8](#_Toc140786767)

[Association Rule Approach 8](#_Toc140786768)

[Apriori 8](#_Toc140786769)

[Fpgrowth 8](#_Toc140786770)

[Model Development 9](#_Toc140786771)

[Predicting Caravan Customer 9](#_Toc140786772)

[Classification Methods (NB, DT, RF, LR, KNN) 9](#_Toc140786773)

[Describe Caravan Customer 9](#_Toc140786774)

[Association Rule Methods (Apriori, Fpgrowth) 9](#_Toc140786775)

[Evaluation and Results 10](#_Toc140786776)

[Validation / Cross-validation / Performance Metrics 10](#_Toc140786777)

[Results and Recommendations 10](#_Toc140786778)

[Conclusion and Further Work 11](#_Toc140786779)

[Limitations and Challenges 11](#_Toc140786780)

[Further Development 11](#_Toc140786781)

[Appendices and References 12](#_Toc140786782)

[References 12](#_Toc140786783)

[Data Dictionary 13](#_Toc140786784)

[Technology use 17](#_Toc140786785)

# 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?

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 selected as 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.

# Introduction

## Context and Objective

A Norwegian insurance company was interested in a machine learning solution to find best customers to market its caravan insurance product. This analysis is to provide recommendations to cross sell their product; caravan insurance. Without sending mass email to all customers, it is cost effective for the company to identify best possible customers who will buy caravan insurance and only approach those customers.

## Approach

This research is completed in these steps.

**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 – Model Development**  
Predicting Caravan Customer - Classification Methods (NB, DT, RF, LR, KNN)

Caravan Customer Description - Association Rule Methods (Apriori, Fpgrowth)

**Step4 - Evaluate the Model**

Validation / Cross-validation / Performance Metrics

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.

# Literature Review

## Background

The data was supplied by [Sentient Machine Research](http://www.smr.nl/). (url: <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.

## Recent Studies

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](https://beginanalyticsblog.wordpress.com/author/beginanalyticsblog/) 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.

# Data Collection and Analysis

## Data Source

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

## Exploratory Data Analysis

Distinct profiles of Caravan Insurance customers based on sociodemographic data

Distinct profiles of Caravan Insurance customers based on policy data

# Data Preprocessing and Preparation

## Data Cleaning

## Feature Engineering / Selection

# Methodology

## Predictive modelling approach

### Multinomial Naïve Bayes

### Decision Tree

### Random forest

### Logistic regression

### K Nearest Neighbours

## Association Rule Approach

### Apriori

### Fpgrowth

# Model Development

### Predicting Caravan Customer

### Classification Methods (NB, DT, RF, LR, KNN)

* predicting a customer's likelihood to purchase Caravan Insurance based on their sociodemographic characteristics

### Describe Caravan Customer

### Association Rule Methods (Apriori, Fpgrowth)

* What frequent associations can be identified in the product ownership data?

# Evaluation and Results

## Validation / Cross-validation / Performance Metrics

## Results and Recommendations

# Conclusion and Further Work

## Limitations and Challenges

## Further Development

I will consider following for further development of the model.

* Use additional algorithms that is not used in the analysis
* Prepare some samples to explain the mechanism of each algorithm used in the analysis. Discuss why the particular algorithm giving high or low accuracy for the dataset we analyse.

I will improve the usability of the model by running it on deferent environments.

* Do the complete analysis using SAS, SQL work bench, R
* Run the code in Cloud environments – GCP, Azure, AWS

# Appendices and References

## 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](http://dx.doi.org/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](https://beginanalyticsblog.wordpress.com/author/beginanalyticsblog/). 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

## Data Dictionary

Data set: TICDATA2000.txt

Sociodemographic attributes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Attribute Name** | **English Label** | **Description** | **type** | **mean** | **std** | **min** | **max** |
| 1 | MOSTYPE | sd\_cust\_subtype | Customer Subtype see L0 | cat | 24.25 | 12.85 | 1 | 41 |
| 2 | MAANTHUI | sd\_no\_of\_houses | Number of houses 1 – 10 | num | 1.11 | 0.41 | 1 | 10 |
| 3 | MGEMOMV | sd\_avg\_household | Avg size household 1 – 6 | num | 2.68 | 0.79 | 1 | 5 |
| 4 | MGEMLEEF | sd\_avg\_age\_band | Avg age see L1 | num | 2.99 | 0.81 | 1 | 6 |
| 5 | MOSHOOFD | sd\_cust\_maintype | Customer main type see L2 | cat | 5.77 | 2.86 | 1 | 10 |
| 6 | MGODRK | sd\_religion\_catholic | Roman catholic see L3 | num | 0.70 | 1.00 | 0 | 9 |
| 7 | MGODPR | sd\_religion\_protestant | Protestant ... | num | 4.63 | 1.72 | 0 | 9 |
| 8 | MGODOV | sd\_religion\_other | Other religion | num | 1.07 | 1.02 | 0 | 5 |
| 9 | MGODGE | sd\_religion\_no | No religion | num | 3.26 | 1.60 | 0 | 9 |
| 10 | MRELGE | sd\_rel\_married | Married | num | 6.18 | 1.91 | 0 | 9 |
| 11 | MRELSA | sd\_rel\_living\_tg | Living together | num | 0.88 | 0.97 | 0 | 7 |
| 12 | MRELOV | sd\_rel\_other | Other relation | num | 2.29 | 1.72 | 0 | 9 |
| 13 | MFALLEEN | sd\_rel\_no\_singles | Singles | num | 1.89 | 1.80 | 0 | 9 |
| 14 | MFGEKIND | sd\_hshold\_wo\_children | Household without children | num | 3.23 | 1.62 | 0 | 9 |
| 15 | MFWEKIND | sd\_hshold\_w\_children | Household with children | num | 4.30 | 2.01 | 0 | 9 |
| 16 | MOPLHOOG | sd\_education\_higher | High level education | num | 1.46 | 1.62 | 0 | 9 |
| 17 | MOPLMIDD | sd\_education\_medium | Medium level education | num | 3.35 | 1.76 | 0 | 9 |
| 18 | MOPLLAAG | sd\_education\_lower | Lower level education | num | 1.11 | 0.41 | 1 | 10 |
| 19 | MBERHOOG | sd\_empst\_high | High status | num | 1.90 | 1.80 | 0 | 9 |
| 20 | MBERZELF | sd\_empst\_Entrepr | Entrepreneur | num | 0.40 | 0.78 | 0 | 5 |
| 21 | MBERBOER | sd\_empst\_farmer | Farmer | num | 0.52 | 1.06 | 0 | 9 |
| 22 | MBERMIDD | sd\_empst\_mdl\_mgmt | Middle management | num | 2.90 | 1.84 | 0 | 9 |
| 23 | MBERARBG | sd\_empst\_skill\_labour | Skilled labourers | num | 2.22 | 1.73 | 0 | 9 |
| 24 | MBERARBO | sd\_empst\_unskill\_labour | Unskilled labourers | num | 2.31 | 1.69 | 0 | 9 |
| 25 | MSKA | sd\_socialclassA | Social class A | num | 1.62 | 1.72 | 0 | 9 |
| 26 | MSKB1 | sd\_socialclassB1 | Social class B1 | num | 1.61 | 1.33 | 0 | 9 |
| 27 | MSKB2 | sd\_socialclassB2 | Social class B2 | num | 2.20 | 1.53 | 0 | 9 |
| 28 | MSKC | sd\_socialclassC | Social class C | num | 3.76 | 1.94 | 0 | 9 |
| 29 | MSKD | sd\_socialclassD | Social class D | num | 1.07 | 1.30 | 0 | 9 |
| 30 | MHHUUR | sd\_rentedhouse | Rented house. Rented house, in the zipcode area of the customer | num | 4.24 | 3.09 | 0 | 9 |
| 31 | MHKOOP | sd\_homeowners | Home owners | num | 4.77 | 3.09 | 0 | 9 |

Policy ownership attributes

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Attribute Name** | **English Label** | **Description** | **type** | **mean** | **std** | **min** | **max** |
| 32 | MAUT1 | sd\_car\_1 | 1 car | num | 6.04 | 1.55 | 0 | 9 |
| 33 | MAUT2 | sd\_car\_2 | 2 cars | num | 1.32 | 1.20 | 0 | 7 |
| 34 | MAUT0 | sd\_car\_0 | No car | num | 1.96 | 1.60 | 0 | 9 |
| 35 | MZFONDS | sd\_health\_ins\_national | National Health Service | num | 6.28 | 1.98 | 0 | 9 |
| 36 | MZPART | sd\_health\_ins\_private | Private health insurance | num | 2.73 | 1.98 | 0 | 9 |
| 37 | MINKM30 | sd\_income\_l\_30k | Income < 30 | num | 2.57 | 2.09 | 0 | 9 |
| 38 | MINK3045 | sd\_income\_30k\_45k | Income 30-45.000 | num | 3.54 | 1.88 | 0 | 9 |
| 39 | MINK4575 | sd\_income\_45k\_75k | Income 45-75.000 | num | 2.73 | 1.93 | 0 | 9 |
| 40 | MINK7512 | sd\_income\_75k\_122k | Income 75-122.000 | num | 0.80 | 1.16 | 0 | 9 |
| 41 | MINK123M | sd\_income\_g\_123k | Income >123.000 | num | 0.20 | 0.55 | 0 | 9 |
| 42 | MINKGEM | sd\_income\_avg | Average income  % of people having average income | num | 3.78 | 1.32 | 0 | 9 |
| 43 | MKOOPKLA | sd\_p\_power\_class | Purchasing power class | num | 4.24 | 2.01 | 1 | 8 |
| 44 | PWAPART | po\_ins\_pol\_thirdparty\_pvt | Contribution private third party insurance see L4 | num | 0.77 | 0.96 | 0 | 3 |
| 45 | PWABEDR | po\_ins\_pol\_thirdparty\_firms | Contribution third party insurance (firms) ... | num | 0.04 | 0.36 | 0 | 6 |
| 46 | PWALAND | po\_ins\_pol\_thirdparty\_agri | Contribution third party insurane (agriculture) | num | 0.07 | 0.50 | 0 | 4 |
| 47 | PPERSAUT | po\_ins\_pol\_car | Contribution car policies | num | 2.97 | 2.92 | 0 | 8 |
| 48 | PBESAUT | po\_ins\_pol\_del\_van | Contribution delivery van policies | num | 0.05 | 0.53 | 0 | 7 |
| 49 | PMOTSCO | po\_ins\_pol\_motorcycle\_sc | Contribution motorcycle/scooter policies | num | 0.18 | 0.90 | 0 | 7 |
| 50 | PVRAAUT | po\_ins\_pol\_lorry | Contribution lorry policies | num | 0.01 | 0.24 | 0 | 9 |
| 51 | PAANHANG | po\_ins\_pol\_trailer | Contribution trailer policies | num | 0.02 | 0.21 | 0 | 5 |
| 52 | PTRACTOR | po\_ins\_pol\_tractor | Contribution tractor policies | num | 0.09 | 0.60 | 0 | 6 |
| 53 | PWERKT | po\_ins\_pol\_agri\_machines | Contribution agricultural machines policies | num | 0.01 | 0.23 | 0 | 6 |
| 54 | PBROM | po\_ins\_pol\_moped | Contribution moped policies | num | 0.22 | 0.81 | 0 | 6 |
| 55 | PLEVEN | po\_ins\_pol\_life | Contribution life insurances | num | 0.19 | 0.90 | 0 | 9 |
| 56 | PPERSONG | po\_ins\_pol\_accident\_ins\_pvt | Contribution private accident insurance policies | num | 0.01 | 0.21 | 0 | 6 |
| 57 | PGEZONG | po\_ins\_pol\_accident\_ins\_fam | Contribution family accidents insurance policies | num | 0.02 | 0.19 | 0 | 3 |
| 58 | PWAOREG | po\_ins\_pol\_disability | Contribution disability insurance policies | num | 0.02 | 0.38 | 0 | 7 |
| 59 | PBRAND | po\_ins\_pol\_fire | Contribution fire policies | num | 1.83 | 1.88 | 0 | 8 |
| 60 | PZEILPL | po\_ins\_pol\_surfboard | Contribution surfboard policies | num | 0.00 | 0.04 | 0 | 3 |
| 61 | PPLEZIER | po\_ins\_pol\_boat | Contribution boat policies | num | 0.02 | 0.27 | 0 | 6 |
| 62 | PFIETS | po\_ins\_pol\_bicycle | Contribution bicycle policies | num | 0.03 | 0.16 | 0 | 1 |
| 63 | PINBOED | po\_ins\_pol\_property | Contribution property insurance policies | num | 0.02 | 0.20 | 0 | 6 |
| 64 | PBYSTAND | po\_ins\_pol\_social security | Contribution social security insurance policies | num | 0.05 | 0.41 | 0 | 5 |
| 65 | AWAPART | po\_no\_ins\_pol\_thirdparty\_pvt | Number of private third party insurance | num | 0.40 | 0.49 | 0 | 2 |
| 66 | AWABEDR | po\_no\_ins\_pol\_thirdparty\_firms | Number of third party insurance (firms) | num | 0.01 | 0.13 | 0 | 5 |
| 67 | AWALAND | po\_no\_ins\_pol\_thirdparty\_agri | Number of third party insurane (agriculture) | num | 0.02 | 0.14 | 0 | 1 |
| 68 | APERSAUT | po\_no\_ins\_pol\_car | Number of car policies | num | 0.56 | 0.60 | 0 | 7 |
| 69 | ABESAUT | po\_no\_ins\_pol\_del\_van | Number of delivery van policies | num | 0.01 | 0.13 | 0 | 4 |
| 70 | AMOTSCO | po\_no\_ins\_pol\_motorcycle\_sc | Number of motorcycle/scooter policies | num | 0.04 | 0.23 | 0 | 8 |
| 71 | AVRAAUT | po\_no\_ins\_pol\_lorry | Number of lorry policies | num | 0.00 | 0.06 | 0 | 3 |
| 72 | AAANHANG | po\_no\_ins\_pol\_trailer | Number of trailer policies | num | 0.01 | 0.13 | 0 | 3 |
| 73 | ATRACTOR | po\_no\_ins\_pol\_tractor | Number of tractor policies | num | 0.03 | 0.24 | 0 | 4 |
| 74 | AWERKT | po\_no\_ins\_pol\_agri\_machines | Number of agricultural machines policies | num | 0.01 | 0.12 | 0 | 6 |
| 75 | ABROM | po\_no\_ins\_pol\_moped | Number of moped policies | num | 0.07 | 0.27 | 0 | 2 |
| 76 | ALEVEN | po\_no\_ins\_pol\_life | Number of life insurances | num | 0.08 | 0.38 | 0 | 8 |
| 77 | APERSONG | po\_no\_ins\_pol\_accident\_ins\_pvt | Number of private accident insurance policies | num | 0.01 | 0.07 | 0 | 1 |
| 78 | AGEZONG | po\_no\_ins\_pol\_accident\_ins\_fam | Number of family accidents insurance policies | num | 0.01 | 0.08 | 0 | 1 |
| 79 | AWAOREG | po\_no\_ins\_pol\_disability | Number of disability insurance policies | num | 0.00 | 0.08 | 0 | 2 |
| 80 | ABRAND | po\_no\_ins\_pol\_fire | Number of fire policies | num | 0.57 | 0.56 | 0 | 7 |
| 81 | AZEILPL | po\_no\_ins\_pol\_surfboard | Number of surfboard policies | num | 0.00 | 0.02 | 0 | 1 |
| 82 | APLEZIER | po\_no\_ins\_pol\_boat | Number of boat policies | num | 0.01 | 0.08 | 0 | 2 |
| 83 | AFIETS | po\_no\_ins\_pol\_bicycle | Number of bicycle policies | num | 0.03 | 0.21 | 0 | 3 |
| 84 | AINBOED | po\_no\_ins\_pol\_property | Number of property insurance policies | num | 0.01 | 0.09 | 0 | 2 |
| 85 | ABYSTAND | po\_no\_ins\_pol\_social security | Number of social security insurance policies | num | 0.01 | 0.12 | 0 | 2 |
| 86 | CARAVAN | po\_ins\_pol\_caravan | Number of mobile home policies 0 - 1. target variable. | num | 0.06 | 0.24 | 0 | 1 |

## Technology use

|  |  |
| --- | --- |
| Task | Tool/package/ library |
| Data profiling, visualisation, feature engineering | **Python** : pandas, numpy, matplotlib, seaborne, pandas profiling  **SAS**  **Tableau**  **R**: ggplot2, caret |
| Classification model | **Python**: sklearn, python-weka-wrapper3, pandas numpy, matplotlib, graphviz  **R**: ISLR, class, fpc, cluster |
| Association rules | **Python**: weka, numpy, pandas  **R**: ISLR, class, fpc, cluster |

Sample Chart

Sample Table

|  |  |  |  |
| --- | --- | --- | --- |
|  | Header Row | Header Row | Header Row |
| First Column | Item 1 | Item 2 | Item 3 |
| First Column | Item 4 | Item 5 | Item 6 |