Predicting and Explaining Caravan Policy Ownership

CIND820: Capstone Project
Project by: Manohari Wijesooriya
swijesooriya@torontomu.ca
Student #501212269

Ryerson University

Supervisor: Dr. Tamer Abdou
Date of Submission: July 20, 2023

Table of Contents

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. (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.

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 Bayes algorithm in Python in predicting caravan customers.

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.

Policy count variables are a numeric value representing policy count. Target variable represent existence of caravan policy with 0 / 1. All other variables are scaled with numeric value. Average age 3 represents 30 – 40 years.

The attribute labels in the file were relabelled with English names for easy of reference. Eg. MHKOOP is relabeled as sd_homeowners.

See appendix for the data dictionary.

Exploratory Data Analysis

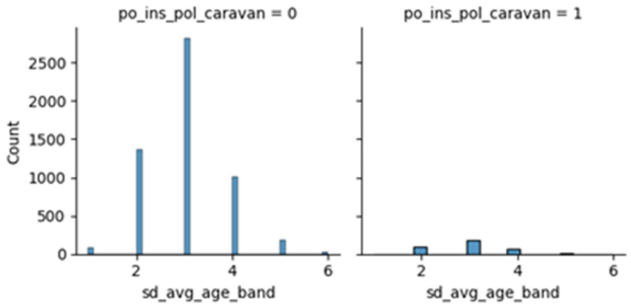
This is an imbalanced dataset. Records with Caravan insurance = 1 is only 6.5% of the dataset.

The training set contains 5474 0’s and 348 1’s. The test set contains 3762 0’s and 238 1’s.

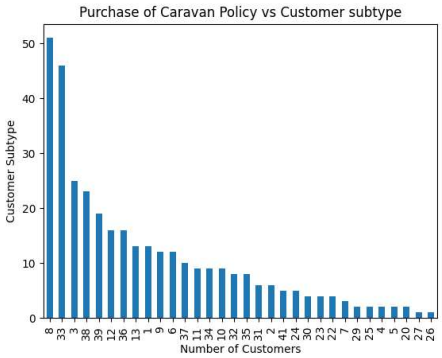
Variable 1-43 are socio-demographic data and variables 44-86 are product ownership data. The socio-demographic data is given by zip code. All customers living in the same zip code have the same sociodemographic attributes.

Distinct profiles of Caravan Insurance customers based on sociodemographic data

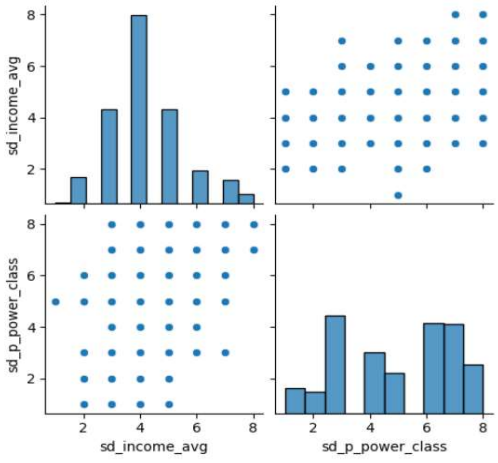
Most customers are in the age range of 40 – 50 years



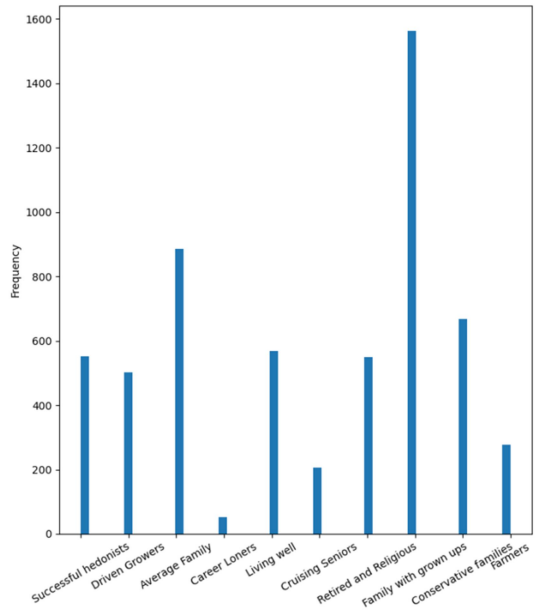
Most customers are middle class families or lower class large families



High average income for higher purchasing power class

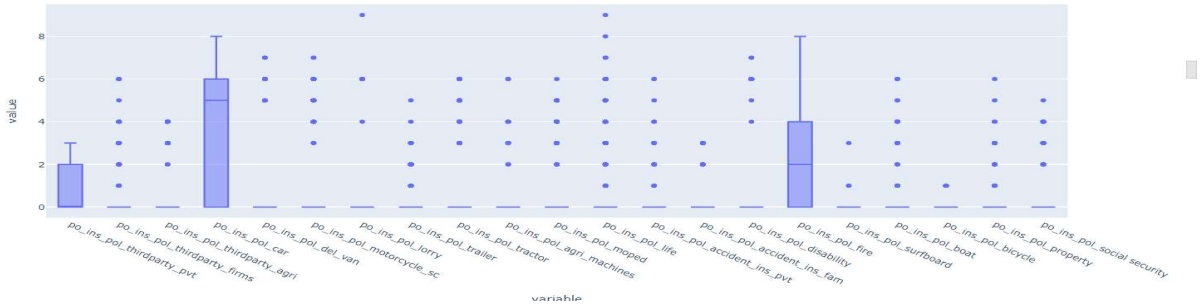


Most of customers are retired and religeous families



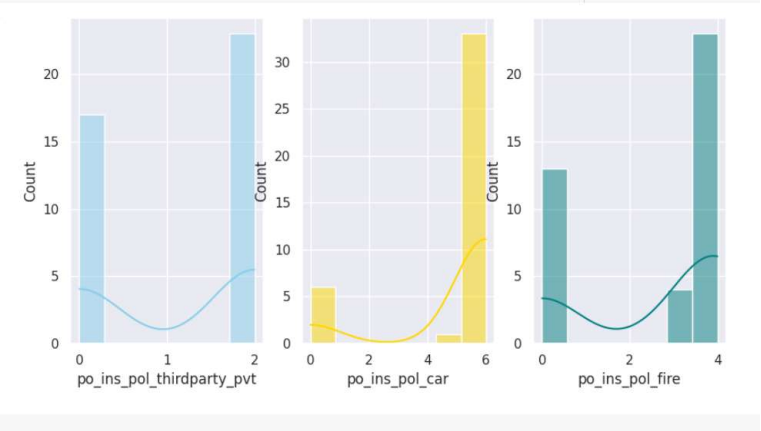
Distinct profiles of Caravan Insurance customers based on Product data

Popular products – third party private, car and fire policy



Popular products – count and percentage

Product	Count	Percentage (out of 5822 customers)
Fire	2270	39%
Car	2150	37%
Third Party Private	1749	30%
Scooter	294	5%

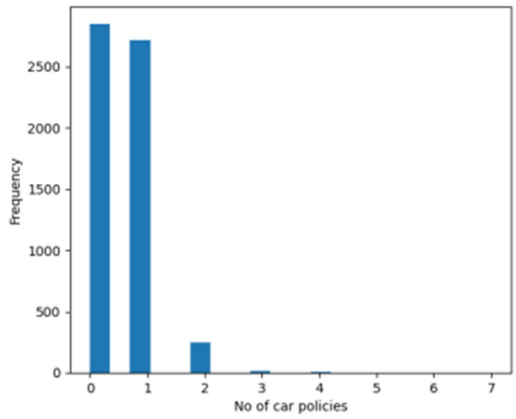


Lower third party insurance premiums

High car insurance premiums

High fire insurance premiums

Most customers have either 1 car or no car



Data Preprocessing and Preparation

Data Cleaning

There are no null values in the dataset. There are 602 duplicate records, but were not removed as sociodemographic variables are same for customers from the same postal code and contribution to the policy attributes are scaled. So same customer may have same values in the dataset, even actual numbers may different. Hence duplicate records were not removed.

Some constraints were checked to understand the validity of data.

- Policy count and contribution to the policy variables have consistent values
- Are there records without product details?
- Check income level variables customers with higher number of policies or higher contribution to policy.

All these tests were passed with good results confirming data integrity.

Feature Engineering / Selection

There were many attributes with low variance (surf board policies, lorry policy). Low variance variables were identified with a threshold of 0.1 (when there is an attribute with same value more than 99%). At the same time, customers buying habit was examined. Only 0.22% of customers have a family accident insurance policy, out of those customers, 40% have a caravan policy. Even our low variance test listed accident and boat policy to drop, these 2 were kept in the dataset to analyse. This reduced number of attributes to 65.

Attributes that has strong correlation to another variable are removed. Customer main type and customer sub type are strongly correlated. These 2 are correlated to purchasing power class. Fire policy has strong correlation with third party private policy. Dropping high correlated features reduced number of attributes to 56.

Some new variables were derived combining attributes.

1) ‘dr_no_car_tptypvt’ : 1 if there is a car or fire policy, 0- no car or third party policy
2) ‘dr_car_tptypvt’: 0 – no contribution to car or third party policy, 1 – one or both car/ third party policy contribution in 1-5 category, 2 - one or both car/ third party policy contribution in 6-9 category
3) ‘dr_no_boat_famacc’ – 0 – no boat or family accident insurance policy, 1 – having boat or family accident insurance policy.

Policy count variables were kept as numeric and all other attributes were changed to string type.

Applied one hot encoding to (converted) categorical variables. This increased the number of columns to 426.

Dataset is split to training and test sets with ratio of 70/30 validation/test sets.

As the dataset has low number of target =1 records, Synthetic Minority Oversampling Technique (SMOTE) is used to balance the training set.

Training Data	Target = 0	Target = 1
Before	3833	242
With SMOTE	3833	3833

Models NB, DT, RF, LR, KNN were trained with encoded training data and evaluated results examining model accuracy, precision and recall.

Then n numbers of features are selected as the features with best chi-square.

Models were run with selected features and reviewed results. This is done iteratively until we find the best selection of features to give highest true positives.

Methodology

Predictive modelling approach

Multinomial Naïve Bayes – NB

This is a probabilistic model which assumes conditional independence between features.

Dependent probability is based on what is the chance of some outcome given some other outcome.

Decision Tree – DT

A decision tree is a graph that makes use of branching method to demonstrate every possible outcome of a decision. In classification, the data is segregated based on a series of questions. Decision Tree has a faster learning speed than other classification methods. This can be converted to easy and simple classification rules.

Random Forest – RF

A random forest can be considered an ensemble of decision trees. It builds and combines multiple decision trees to get a more accurate prediction. Each of the decision tree models used is weak when employed on its own, but it becomes stable when put together.

Logistic Regression – LR

This method is widely used for binary classification problems. Here the dependent variable is categorical and have only 2 values, like 0 or 1, win or lose or pass or fail etc. Here model is detecting maximum likelihood of something happening.

K Nearest Neighbours – KNN

This is an algorithm that classifies data points by a majority vote of its k neighbors. It is used to assign a data point to clusters based on similarity measurement. A new input point is classified in the category such that it has the most number from that category. For example Making an email as spam or ham.

Association Rule Approach

Association rule mining is a rule-based machine learning technique used to find frequent patterns in a data set. Frequent patterns may include frequent item sets that are usually bought together or subsequences that are bought in sequence.

So, in a given transaction with multiple items, Association Rule Mining primarily tries to find the rules that govern how or why such products/items are often bought together.

Apriori

Apriori is a join based algorithm for frequent item set mining and association rule learning over a sample set. It proceeds by identifying the frequent individual items in the sample and extending

them to larger and larger item sets as long as those item sets appear sufficiently often in the sample.

Fpgrowth

Frequent Pattern Growth (FP-Growth) is one of the alternative algorithms that can be used to determine the most common set of data in a dataset. The algorithm searches for the Association Rule by using the value parameters of support and confidence.

Model Evaluation Methodology

Classification models -

Parameters of confusion matrix are helpful in evaluating performance of various models.

Confusion Matrix

A confusion matrix examines all possible outcomes of prediction: true positive, true negative, false positive and false negative.

The parameters calculated from a confusion matrix are:

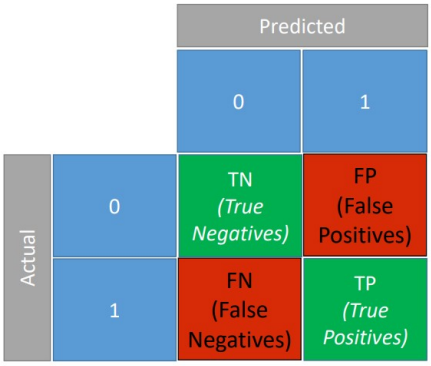
• Accuracy rate: The proportion of the total number of predictions that was right

• Precision/Positive Predicted Value: The proportion of positive cases that were correctly identified

• Negative Predictive Value: The proportion of negative cases that were correctly identified

• Recall/Sensitivity/True Positive Rate: The proportion of actual positive cases which are correctly identified

• Specificity/ True Negative Rate: The proportion of actual negative cases which are correctly identified



Accuracy Rate = *(TP + TN) /(TP+TN+FP+FN)*

Error Rate = *(FP + FN) /(TP+TN+FP+FN)*

Precision/ Positive Predicted Value = *(TP) /(TP+FP)*

Recall/Sensitivity/True Positive Rate = *(TP) /(TP+FN)*

Specificity/ True Negative Rate= *(TN) /(TN+FP)*

Association rule -

Metrics for evaluating association rules and setting selection thresholds are listed below. Given a rule "A -> C", A stands for antecedent and C stands for consequent.

SUPPORT = A simple way to control complexity is to place a constraint that such rules must apply to some minimum percentage of the data

$$\text{support}\,(A \rightarrow C) = \text{support}(A \cup C), \quad \text{range: } [0,1]$$

CONFIDENCE = The probability that B occurs when A; it is p(B|A), which in association mining.

$$\text{confidence}\,(A \rightarrow C) = \text{support}(A \rightarrow C)\,\text{support}(A), \quad \text{range: } [0,1]$$

LIFT = the co-occurrence of A and B is the probability that we actually see the two together, compared to the probability that we would see the two together if they were unrelated to (independent of) each other.

$$\text{lift}\,(A \rightarrow C) = \text{confidence}(A \rightarrow C)\,\text{support}(C), \quad \text{range: } [0,\infty]$$

LEVERAGE = alternative is to look at the difference between these quantities rather than their ratio.

$$\text{leverage}\,(A \rightarrow C) = \text{support}(A \rightarrow C) - \text{support}(A) \times \text{support}(C), \quad \text{range: } [-1,1]$$

CONVICTION = measure to ascertain the direction of the rule. Unlike lift, conviction is sensitive to the rule direction.

$$\text{conviction}\,(A \rightarrow C) = 1 - \text{support}(C)\,1 - \text{confidence}(A \rightarrow C), \quad \text{range: } [0,\infty]$$

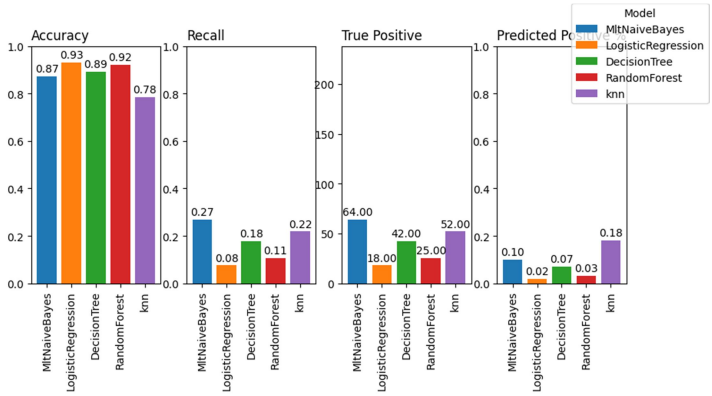
Model Development

Predicting Caravan Customer

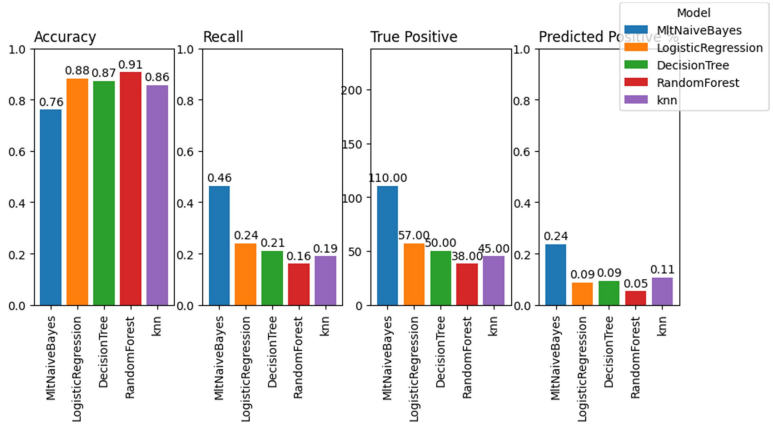
- predicting a customer's likelihood to purchase Caravan Insurance based on their **sociodemographic and product** characteristics

After data cleaning, modelling steps started with 59 attributes. Models NB, DT, RF, LR, KNN were trained with encoded training data and evaluated results examining model accuracy, precision and recall.

- With 59 training variables (426 encoded columns) –
  NB predict 64 true positive counts with accuracy of 73%

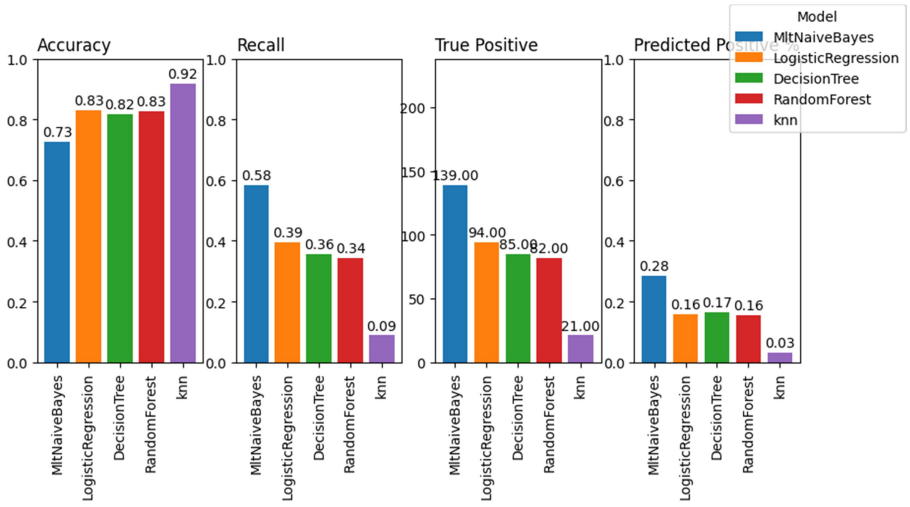


- With 80 encoded columns –
  NB predict 110 true positive counts with accuracy of 76%



- With 28 encoded columns –

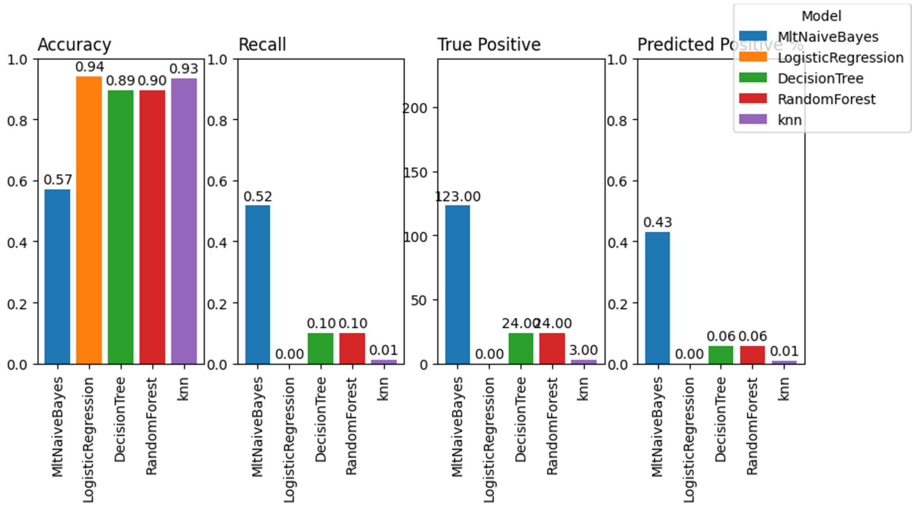
  NB predict 139 true positive counts with accuracy of 73%



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

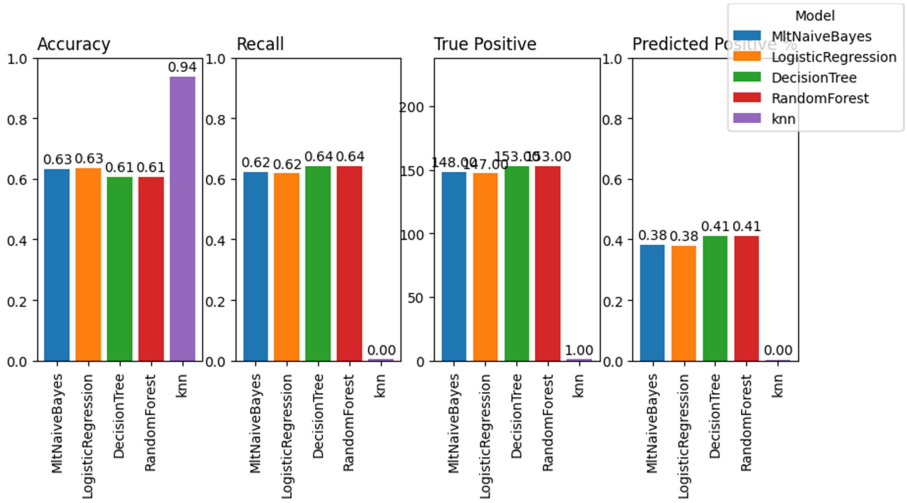
Models were run by keeping only sociodemographic data. Using Naïve Bayes algorithm,

We could predict 123 true positive count with 57% accuracy.



- predicting a customer's likelihood to purchase Caravan Insurance based on their product characteristics

Models were run keeping product data. Using 23 product attributes converted to 85 encoded attributes NB model predicted 138 true positive targets with 64% accuracy. By using best Chi squared values 12 attributes were selected. This gave the best model results by predicting 148 true positive counts with 64% accuracy. Higher number of true positive was predicted using only product variables than using with sociodemographic and product data.



Using all data we can build a model with high accuracy than using sociodemographic and product data separately.

Caravan Customer Description

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

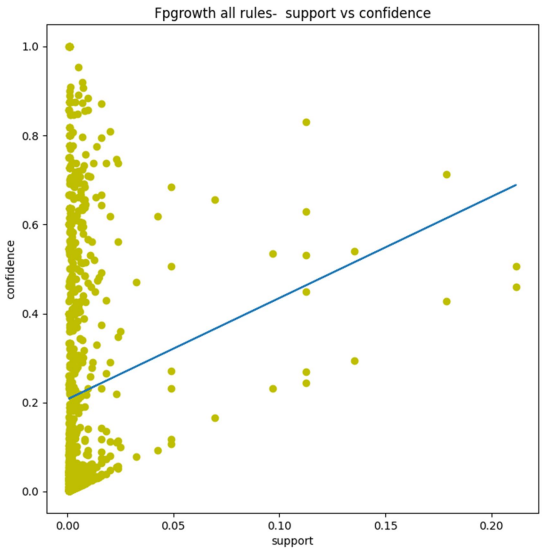
Fist frequent associations were tested using **product premium** variables. Following are results from initial baseline models with sample 8000 records.

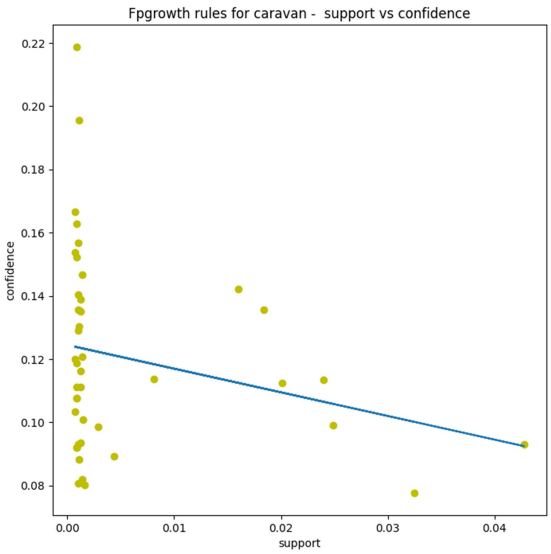
Other products that best associate with caravan policy:

	Apriori	Fpgrowth
Model parameters	Frequent items with support 0.01% metric="lift", max_len = 4 min_threshold= 1	Frequent items with support 0.01% metric="lift", max_len = 4 min_threshold= 1
5 recommendations with caravan product	{'po_ins_pol_bicycle_1', 'po_ins_pol_car_6'} {'po_ins_pol_car_6', 'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2'} {'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2', 'po_ins_pol_motorcycle_sc_4'} {'po_ins_pol_moped_3', 'po_ins_pol_car_6'} {'po_ins_pol_car_6', 'po_ins_pol_fire_4'}	{'po_ins_pol_bicycle_1', 'po_ins_pol_car_6'} {'po_ins_pol_car_6', 'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2'} {'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2', 'po_ins_pol_motorcycle_sc_4'} {'po_ins_pol_moped_3', 'po_ins_pol_car_6'} {'po_ins_pol_car_6', 'po_ins_pol_fire_4'}

Fpgrowth algorithm results can be explained as this. Customers who buy a bicycle policy in premium category 1 (1 – 49%) and a car policy in premium category 6 (1000 – 4999) will also choose a caravan policy.

Movement of Lift and Confidence of Fpgrowth rules





Confidence takes values from 0 to 1. Lift can be from 0 to infinity. Rules with high confidence are favorable.

Evaluation and Results

Validation / Cross-validation / Performance Metrics

Classification model (predicting /identifying caravan customer)

The aim of this research is to identify caravan customers from given attributes. We are trying reducing the cost of marketing team by recommending the best sub set of their customer base to reach to market the Caravan product. Results from various algorithms using changing parameters and different data pipelines are being reviewed using confusion matrix parameters.

The best model will be the model with higher accuracy = a (a > 50%) and higher recall = r (r >50%). Also I looked at the % of cost reduction. Ie. Percentage of positive targets give an idea of the cost reduction.

Five classification models were used with 56 attributes to train the model and predict target in the initial run.

Then n numbers of features are selected as the features with best chi-square. Models were run with selected features and reviewed results. This is done iteratively until we find the best selection of features to give highest true positives.

- Evaluation results from initial run with 59 training variables (426 encoded columns)

Algorithm	Accuracy	Recall on target = 1 / True Positive rate (1)	True positive (2)	% of predicted positive targets (3)
NB	87%	27%	64	10%
DT	89%	18%	42	2%
RF	92%	11%	25	7%
LR	93%	8%	18	3%
KNN(3)	78%	22%	52	18%

(1) True Positive rate = True Positive / (True Positive + False Negative)
(2) Out of actual positive target = 238
(3) % of predicted positive targets = (True positive + False positive) / number of records in the test sample. This is an indication of cost reduction for the business.

- Evaluation results 80 encoded training variables

Algorithm	Accuracy	Recall on positive target	True positive count	% of predicted positive targets (1)
NB	76%	46%	110	24%
DT	87%	21%	50	9%
RF	91%	16%	38	9%
LR	88%	24%	57	5%
KNN(3)	86%	19%	45	11%

- Evaluation results 28 encoded training variables

Algorithm	Accuracy	Recall on positive target	True positive count	% of predicted positive targets (1)
NB	73%	58%	139	28%
DT	82%	36%	85	16%
RF	83%	34%	82	17%
LR	83%	39%	94	16%
KNN(3)	92%	9%	21	3%

- Evaluation results with 12 encoded product variables

Algorithm	Accuracy	Recall on positive target	True positive count	% of predicted positive targets (1)
NB	63%	62%	148	38%
DT	61%	64%	153	41%
RF	61%	64%	153	41%
LR	63%	62%	147	38%
KNN(3)	94%	0%	21	0%

Association Rules (describe caravan customer)

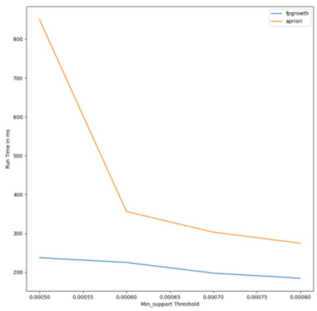
There are three common ways to measure association. That is using support, confidence and lift of the rules.

Criteria used for selecting rules:

Antecedent support > 0.007, confidence > 0.115, lift > 1

One of the most important features of any frequent item set mining algorithm is that it should take lower timing and memory.

Best algorithm is selected based on the run time of 2 models.



Fpgrowth algorithm gives results faster.

Results and Recommendations

Classification model (predicting /identifying caravan customer)

This is the model that could predict most of caravan customers.

Naïve Bayes - train validation split 70% - 30%, using 28 one-hot encoded attributes.

Classification report:

```
                precision    recall  f1-score   support

           0       0.97      0.73      0.83      3762
           1       0.12      0.58      0.20       238

    accuracy                           0.73      4000
   macro avg       0.54      0.66      0.52      4000
weighted avg       0.92      0.73      0.80      4000

   confusion matrix
 [[2765  997]
 [  99  139]]
TP:  139 , FP:  997 , TN:  2765 , FN: 99
accuracy  0.726
recall    0.584
```

The marketing team of the insurance company can use predictions to reach out to 28% of their customer base and get 139 caravan insurance policies.

Association Rules (describe caravan customer)

It is recommended to use Fpgrowth algorithm as the model runs in less run time compared to Apriori algorithm.

Best associations of caravan policy with product variables:

Item sets	support	confidence	lift
{'po_ins_pol_bicycle_1', 'po_ins_pol_car_6'}	0.0014	0.1467	2.1218
{'po_ins_pol_car_6', 'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2'}	0.0160	0.1422	2.0575
{'po_ins_pol_fire_4', 'po_ins_pol_thirdparty_pvt_2', 'po_ins_pol_motorcycle_sc_4'}	0.0010	0.1404	2.0304
{'po_ins_pol_moped_3', 'po_ins_pol_car_6'}	0.0013	0.1389	2.0092
{'po_ins_pol_car_6', 'po_ins_pol_fire_4'}	0.0184	0.1357	1.9636

Customers who buy car policy in premium category 6 (1000 - 4999) and bicycle policy in premium category 1 (1 - 49) also buy a caravan policy.

Rules are selected based on highest confidence. We notice those have minimal support.

Rules provide some signals in identifying caravan customer.

Conclusion and Further Work

Limitations and Challenges

It is difficult to get 100% accurate model. While reducing the cost by contacting less number of customers, they may not get possible caravan customers as they didn't contact them.

One challenge of this dataset was having only 6% of positive targets to train the model. This is overcome by using over sampling method.

Further Development

I will consider following for further development of the model.

- Use additional algorithms that is not used in the analysis
- Use different feature selection methods.

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

- Do the complete analysis using SAS, SQL work bench, R
- Prepare more visuals using Tableau
- 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

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] Shamila Nasreen, M A Azamb,.. (2014) Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey

https://core.ac.uk/download/pdf/82306393.pdf Last accessed on July 18, 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