Predicting and Explaining Caravan Policy Ownership

Manohari Wijesooriya
Student #501212269

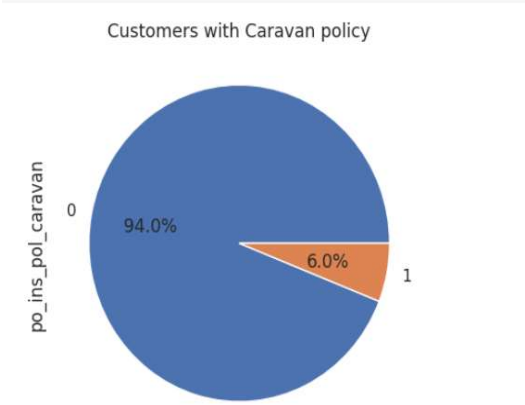
Ryerson University

Introduction

- Insurance company is looking for predictive modelling solution to reduce the cost of marketing of their new product, caravan insurance to internal customers.
- Given sociodemographic and product variables, they are requesting to flag best possible customers to reach out to sell their product.
- Various machine learning algorithms were used in the effort in predicting
- Recommendations – Use Naïve Bayes algorithm for prediction

Data Preparation - Exploration

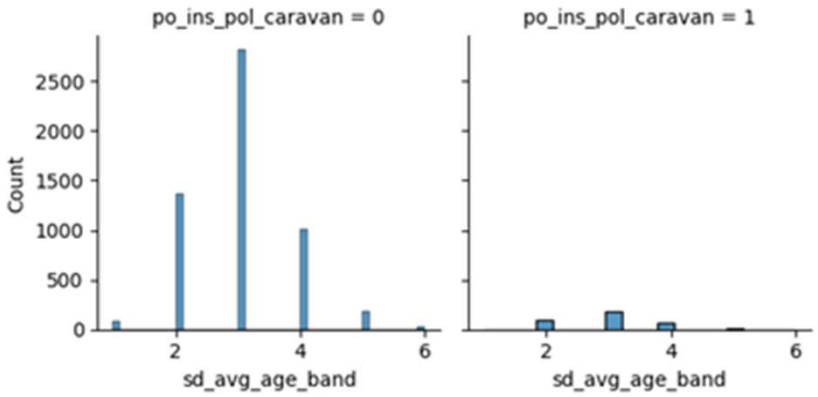
- Explored dataset of 43 socio-demographic and 42 product ownership data
- Target variable : having a caravan insurance policy 0/1 represent 6% of training data
- 2 categorical attributes (customer main / sub type), all others are numeric
- Data cleaning – no null values
- Identify duplicates – 602 records indicated as duplicates but those were not dropped as all sociodemographic data are scaled. Customers in the same postal code have same values.
- Checked integrity of data by running few queries.
- Low variance attributes were identified and removed



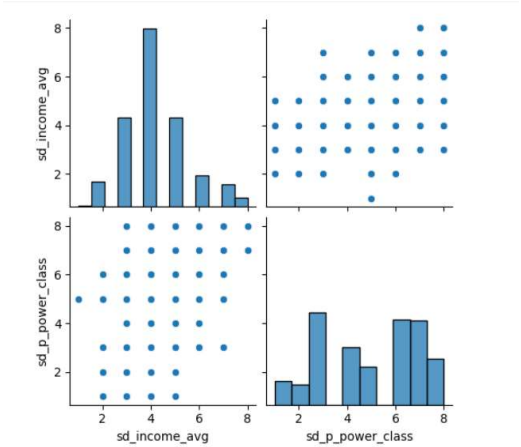
Ryerson University

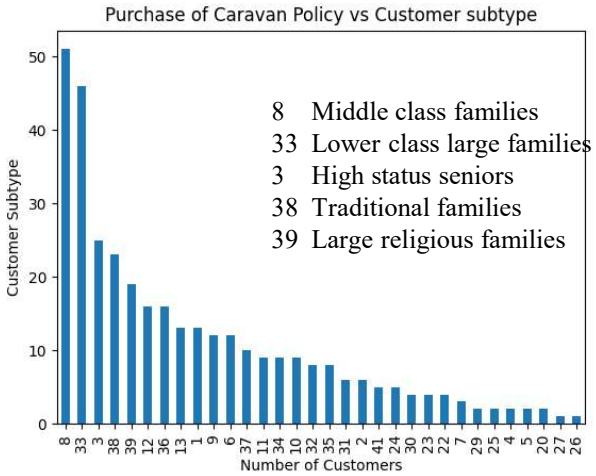
Data Preparation - Exploration

Distinct profiles of Caravan Insurance customers based on sociodemographic data



Most customers are in age range of 40 - 50 years





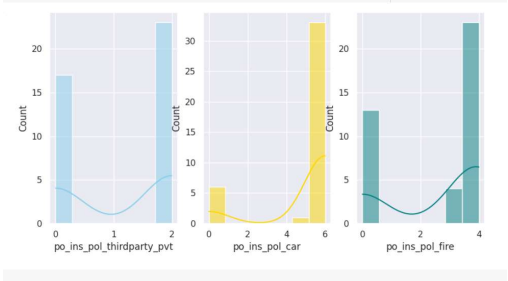
Most customers are middle class families or lower class large families

High average income for higher purchasing power class

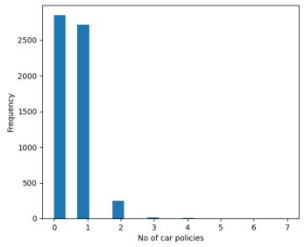
Ryerson University

Data Preparation - Exploration

Distinct profiles of Caravan Insurance customers based on Product data



Lower third party insurance premiums
High car insurance premiums
High fire insurance premiums



Most customers have either 1 car or no car

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

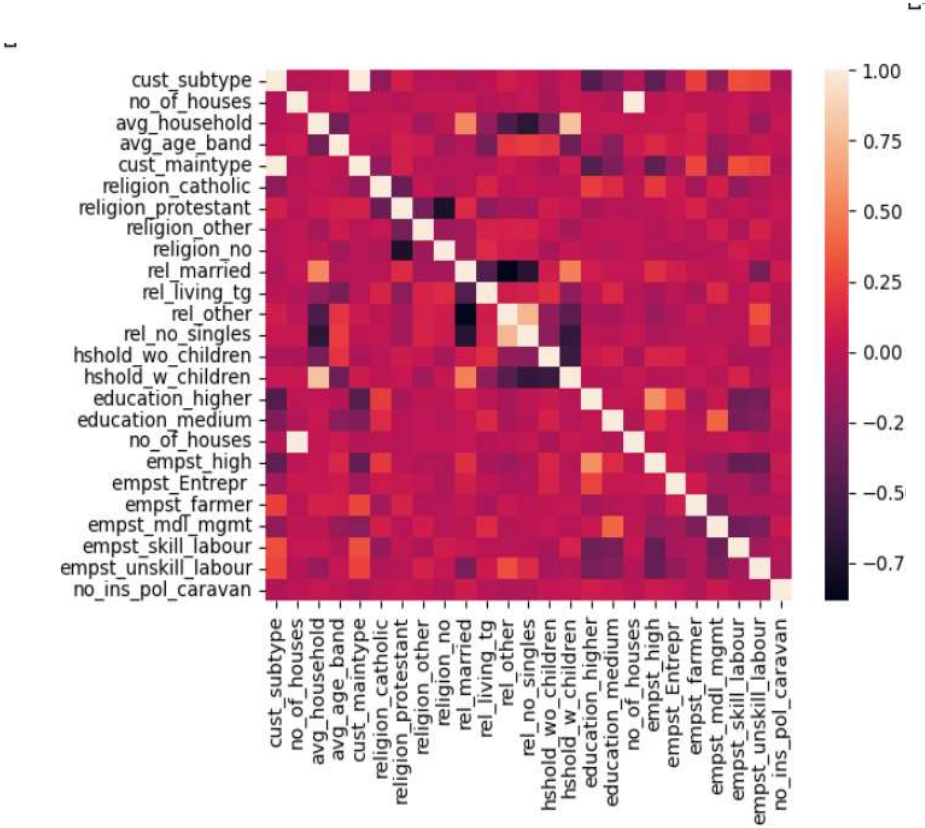
Popular products are third-party private, car and fire policy.

Ryerson University

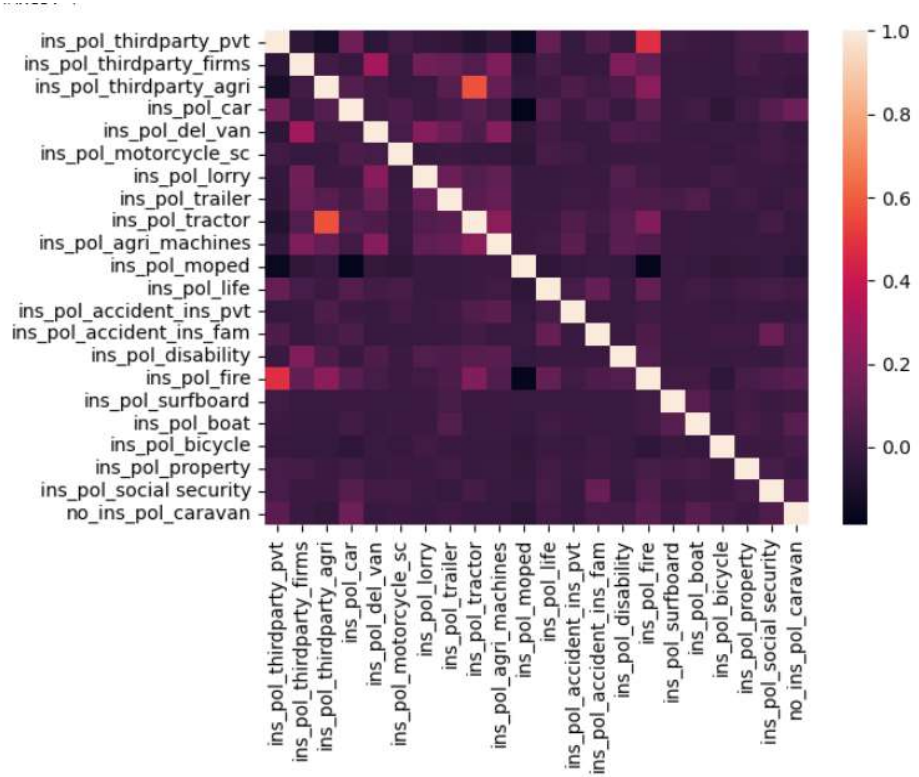
Data Preparation - Exploration

Correlation Matrix

Sociodemographic data



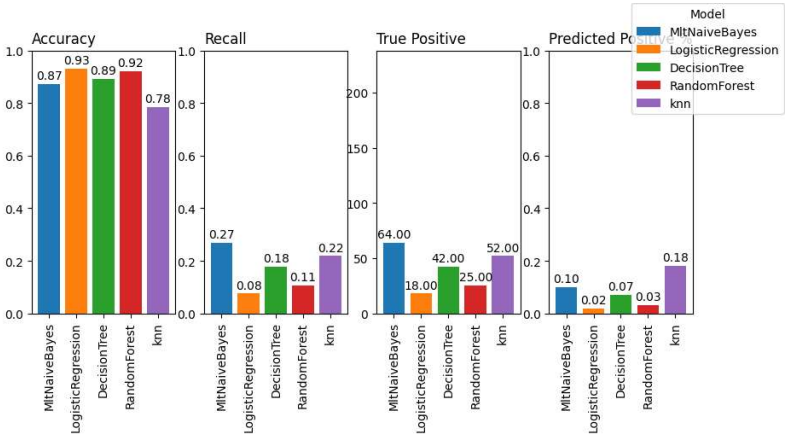
Product data



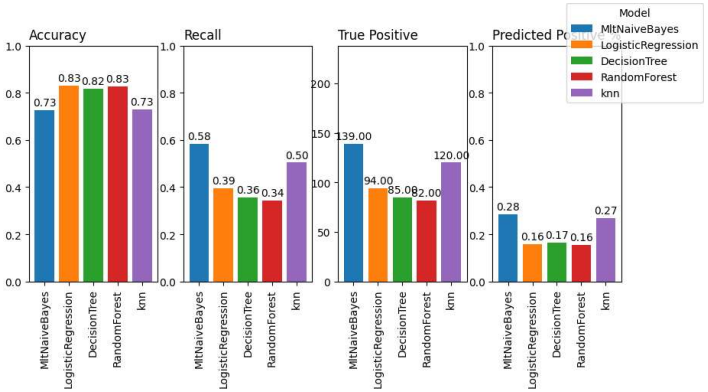
Ryerson University

Predictive Modelling

Baseline model



Final model

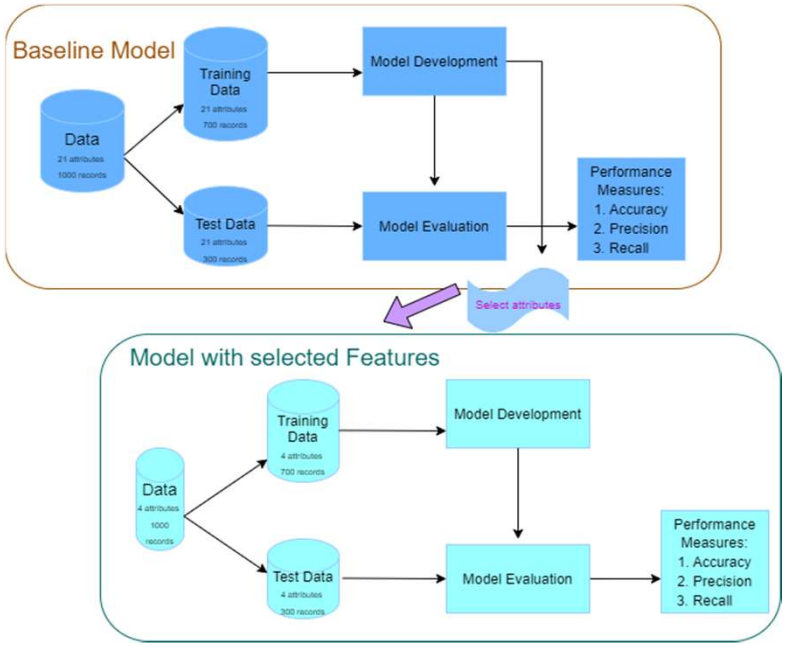


Initial model criteria

Validation/test split : 70/30, random state 7

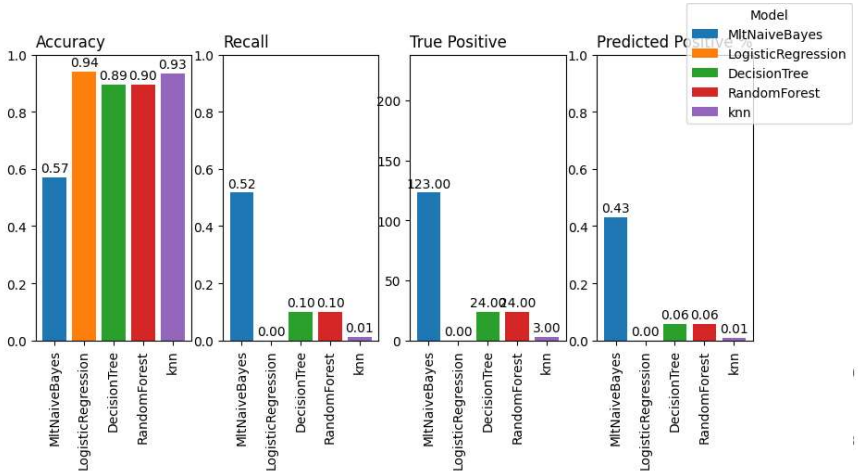
Sampling technique : SMOTE

Encoding categorical attributes



Ryerson University

Predictions by segments of Data



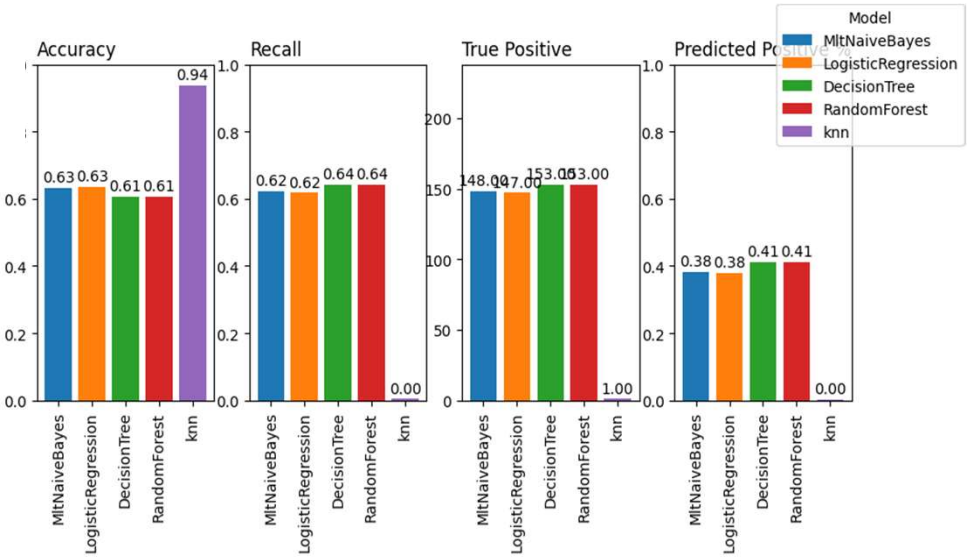
Predicting a customer's likelihood to purchase Caravan Insurance based on their sociodemographic characteristics

Best model - Naïve Bayes
Accuracy 57%, True positive = 123

Best features	
Attribute	Score
sd_religion_other_2	71619.19
sd_empst_unskill_labour_2	62334.3
sd_income_l_30k_2	61913.47
sd_education_medium_3	61581.3
sd_religion_protestant_5	61457.34

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

Best model - Decision Tree , Random Forest
Accuracy 64% True positive = 153

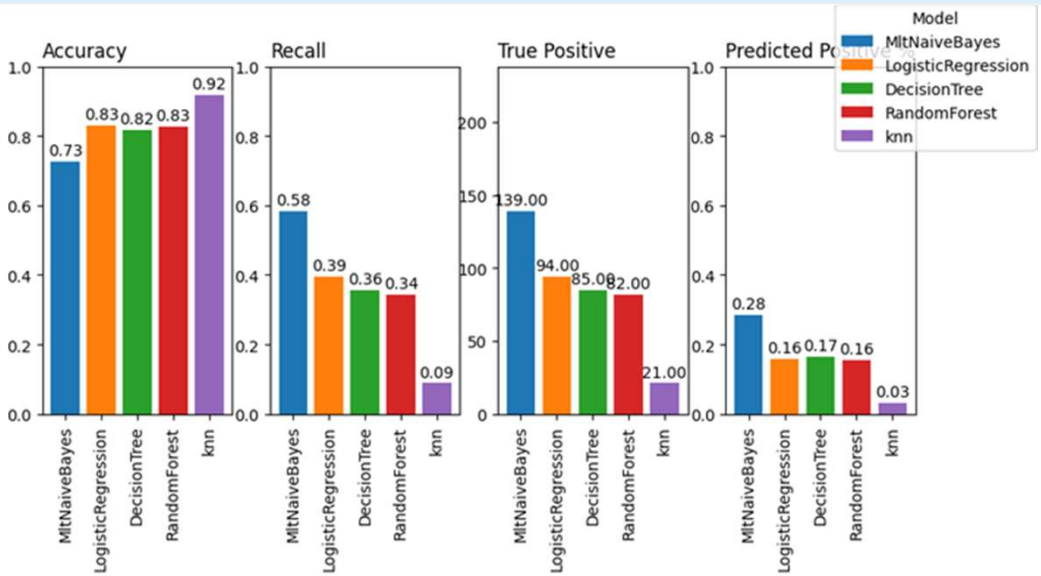


Best features	
Attribute	Score
po_ins_pol_car_6	387.4142
dr_car_tptypvt	367.1447
po_ins_pol_car_0	356.6783
po_no_ins_pol_car	257.3392
po_ins_pol_thirdparty_pvt_2	183.6167

Ryerson University

Predictions by segments of Data

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



Best Attributes	
Attribute	Score
po_ins_pol_car_0	723.2143
dr_car_tptypvt_1	584.9565
dr_car_tptypvt_0	444.7121
po_ins_pol_thirdparty_pvt_0	414.5592
sd_empst_skill_labour_3	404.4648
po_ins_pol_car_5	363.6918
sd_empst_unskill_labour_3	340.1657
sd_income_avg_3	310.5957
sd_socialclassC_5	301.8962
sd_income_l_30k_5	289.1691

Best model - Naïve Bayes
Accuracy 73%, True positive = 139

	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
```

Conclusion and Recommendation

- Naïve Bayes is the best algorithm to predict caravan customers.
- Model could correctly identify 139 caravan customers out of 238 with 73% accuracy
- Model predictions are to contact 28% of the customer base to promote caravan policy
- Best predictive attributes are : car policy, private third party policy, average income 3, skill labour 3, unskill labour 3

Thank you.

Chang School,
Professors,
Technical Assistants,
Colleges

Ryerson University