Machine learning model to predict insurance claims for motor vehicles

1. Introduction

With escalating cost of living worldwide more and more customers are moving to lower cost alternatives which is putting huge pressure on businesses to adopt innovative ways to provide lower cost solutions to the market. This is especially true for the vehicle insurance industry as there is a significant trend for customers to move to lower cost “3rd party only” insurance schemes, moving away from traditional “full cover” insurance policies. The most vulnerable segment to this transition are customers who feel they have a lower risk of insurance claims

Therefore, one method to address this problem is to effectively identify and offer higher discounts or lower more attractive insurance packages to vehicle owners who have a lower risk of initiating insurance claims.

In this context the purpose of the machine learning model is to classify existing and new customers based on the likelihood of requiring an insurance claim within a target period, using customer’s socioeconomic details such as KYC information and other data sources such as police reports available to insurance companies

2. Dataset

A dataset was downloaded from Kaggle named “Car insurance dataset” which has 10000 customer records and 19 columns of data using real customer data and logs of past insurance claims. The data available included below information

FIELD	DESCRIPTION
ID	Unique identification number
AGE	Age in brackets of 16-25 years, 25–39 years, 40-64 years, 65+ years
GENDER	Equal distribution of Male and female
RACE	90% from majority and 10% from minority races
DRIVING_EXPERIENCE	Provided in brackets of 0-9y, 10-19y, 20-29y and 30y+
EDUCATION	Provided in brackets of None, High school and University
INCOME	Provided in brackets of Poverty, Working, middle and upper class
CREDIT_SCORE	Numerical score between 0 and 1
VEHICLE_OWNERSHIP	Boolean 1 or 0
VEHICLE_YEAR	Provided as before 2015 and after 2015
MARRIED	Boolean 1 or 0 indicating maternal status
CHILDREN	Boolean 1 or 0 indicating availability of children
POSTAL_CODE	Numeric code corresponding to address location
ANNUAL_MILEAGE	Numerical value of distance
VEHICLE_TYPE	Classified into Sedan or sports car
SPEEDING_VIOLATIONS	No of occurrences of speeding violations

DUIS	No of occurrences of driving under influence
PAST_ACCIDENTS	No of occurrences of accidents
OUTCOME	Boolean 1 or 0 based on insurance claims requested

3. Methodology

Problem identification

The problem is identified as a classification problem where the objective is to classify customers with a higher likelihood of requesting a claim as Boolean 1 and customers with a lower likelihood of requesting a claim as Boolean 0.

Tools used

Python is selected as the preferred language for developing the model and Google Colab is used as the tool for this development. All data including the input data and corresponding machine learning algorithm is stores in a github repository

Machine learning model selection

Several classifications models available in python are suitable for this type of binary classification problem. Below 4 models are tried out and compared to identify the best model

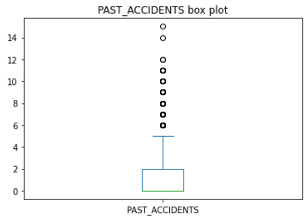
- Logistic regression,
- Decision tree classifier,
- Random forest classifier
- support vector machine.

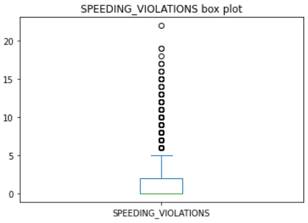
Feature identification

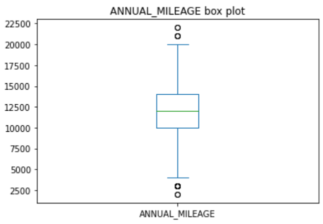
Out of the 19 columns available in the data set "OUTCOME" column is selected as the y variable and remaining 17 columns excluding "ID" are considered as features for analysis.

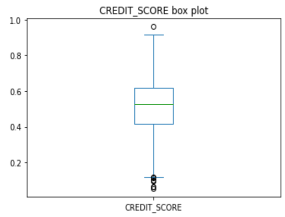
Data pre-processing

During the data pre-processing stage outliers were removed from numerical columns CREDIT_SCORE, ANNUAL_MILEAGE, PAST_ACCIDENTS and SPEEDING_VIOLATIONS. It was observed that DUIS also showed outliers but on further observation since the values were only from 0 to 6 and removal of these outliers negatively impacted the outcome, these were not removed









Missing data was observed in ANNUAL_MILLAGE and CREDIT_SCORE columns and dropping there rows was adopted as treatment

Data encoding was required for categorical columns AGE, GENDER, DRIVING_EXPERIENCE, EDUCATION, INCOME, and VEHICLE_YEAR. Label encoding was followed by replacing the labels with numbers. Encoding was needed POSTAL_CODES as well and label encoding was performed for this as well. The features RACE and VEHICLE_TYPE were not encoded as preliminary analysis showed there was no significant difference in probability between the categories in this features

Training the model

After the pre-processing stage 7464 data points remained in the data set. Out of this a 70:30 split was selected as the train test split, which made 5224 points available as the training data set. All 4 models were trained using the same training data set.

Feature selection

A correlation matrix was used to verify the correlation between the y variable and the 16 features shortlisted. POSTAL_CODE showed the lowest correlation with OUTCOME which was 0.1 while driving experience showed the highest correlation with 0.48.

It was also observed the INCOME and CREDIT_SCORE showed the highest correlation among features at 0.75 while SPEEDING_VIOLATIONS had a 0.64 correlation with DRIVING_EXPERIENCE

Backward elimination method is used for feature selection considering the lower number of samples and the relative low correlation difference compared between features.

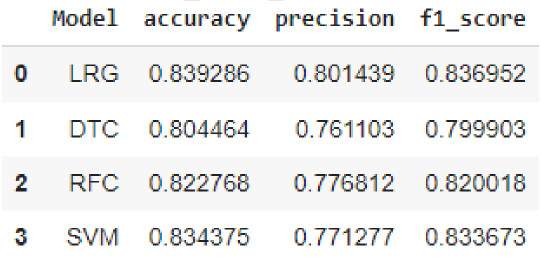
POSTAL_CODE which had the lowest correlation was removed first from feature list but this lower all 3 of Accuracy, Precision and F1 score in all the models. This was added back. Several iterations were carried out eliminating features. It was seen that removing EDUCATION improved the accuracy and F1 score but reduced the pression. This method of eliminating features was carried out iteratively

4. Results

The four models were evaluated using Accuracy, Precision and F1 score. Below are the key results obtained for all the models during iterations

Iteration A - Using 9 features

```
'AGE_NU', 'GENDER_NU', 'DRIVING_EXPERIENCE_NU', 'INCOME_NU', 'CREDIT_SCORE
', 'VEHICLE_OWNERSHIP', 'VEHICLE_YEAR_NU', 'DUIS', 'PAST_ACCIDENTS'
```

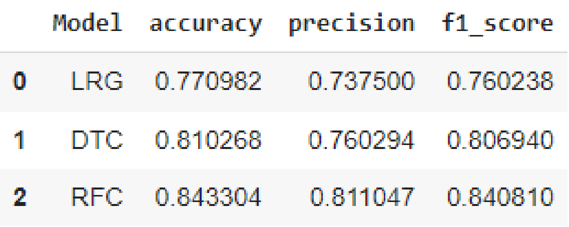


	Model	accuracy	precision	f1_score
0	LRG	0.839286	0.801439	0.836952
1	DTC	0.804464	0.761103	0.799903
2	RFC	0.822768	0.776812	0.820018
3	SVM	0.834375	0.771277	0.833673

The Logistic regression model provided the best outcome with highest accuracy, precision and F1 score

Iteration B – Using 13 features

```
'AGE_NU', 'GENDER_NU', 'DRIVING_EXPERIENCE_NU', 'INCOME_NU', 'CREDIT_SCORE', 'VEHICLE_OWNERSHIP', 'VEHICLE_YEAR_NU', 'MARRIED', 'CHILDREN', 'POSTAL_CODE_NU', 'ANNUAL_MILEAGE', 'DUIS', 'PAST_ACCIDENTS'
```

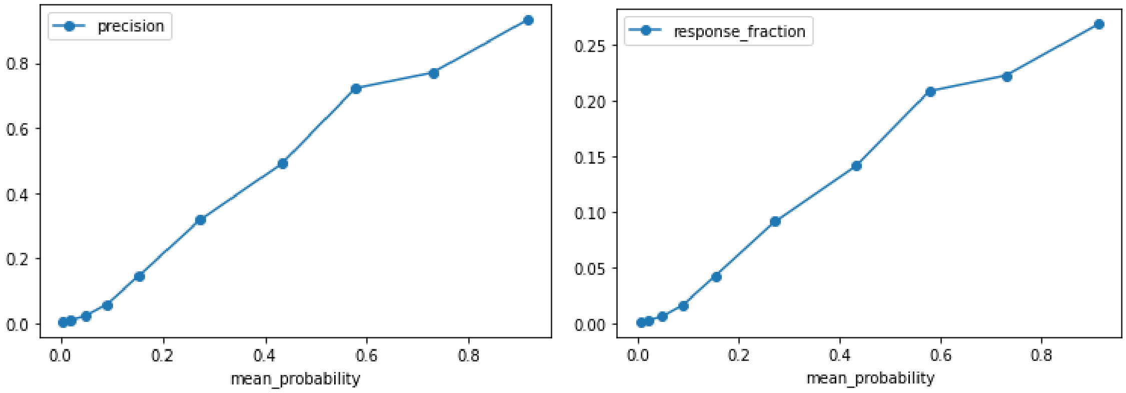


	Model	accuracy	precision	f1_score
0	LRG	0.770982	0.737500	0.760238
1	DTC	0.810268	0.760294	0.806940
2	RFC	0.843304	0.811047	0.840810

The Random-forest classifier gave the best result and the overall best result compared with all iterations and iteration B which is adopted as the final model

It was also seen that the SVM model took an excessively longer time to run with 13 variables and hence was not considered for iteration B.

The following graphs depicts the variation of precision and response-fractions against mean probability for the selected RandomForest classifier model and shows a close to linear behavior.



An attempt was also made at hyper parameter tunning to further increase accuracy, but it was observed that the proposed optimization of 'min_samples_split': 20, 'min_samples_leaf': 10, 'max_depth': 100 did not improve the model further. Hence the original model is continued to be used

5. Conclusion

The proposed Random forest classifier-based model provides an accuracy of 84% and a precision of 81%. This amounts to around 19% of customers getting false positive results which will mean they will not be eligible for a preferential insurance package

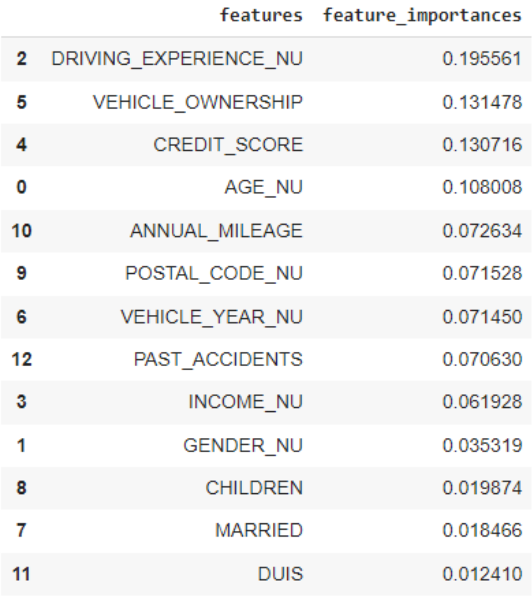
The confusion matrix for the model is as below

		Predicted	
		0	1
Actual	0	1331	130
	1	221	558

Since the objective of the model is to identify low risk customers there are 221 false negatives for a total of 1331 + 221 predicted as negative. This amount to 14% false negative cases which means 14% of users tagged as low risk would need insurance claims

Overall, it can be concluded that the developed model can be effectively used for the purpose of classifying customers based on the likelihood of requesting a car insurance claim

Based on feature importance analysis it can also be seen that driving experience has the highest importance followed by vehicle ownership, credit score and age.



	features	feature_importances
2	DRIVING_EXPERIENCE_NU	0.195561
5	VEHICLE_OWNERSHIP	0.131478
4	CREDIT_SCORE	0.130716
0	AGE_NU	0.108008
10	ANNUAL_MILEAGE	0.072634
9	POSTAL_CODE_NU	0.071528
6	VEHICLE_YEAR_NU	0.071450
12	PAST_ACCIDENTS	0.070630
3	INCOME_NU	0.061928
1	GENDER_NU	0.035319
8	CHILDREN	0.019874
7	MARRIED	0.018466
11	DUIS	0.012410

6. Discussion

The work carried out to develop the model is still at initial stage and further improvements to the model can be done including re-looking at hyper parameter tunning

In addition, the logistic regression model (model 1) in iteration A showed promising results almost similar to the RamdomForest classifier. Further development and optimization of model 1 will be optimum since it needs much less features to achieve this result. This could save storage space and computational power during execution among other advantages.

Further work is also necessary to build a front end for the model