MACHINE LEARNING FOUNDATION JAN 2022
CAPSTONE PROJECT REPORT

Machine learning model to predict insurance claims for motor vehicles

INTRODUCTION

Business problem:

Customers are moving to lower cost “3^{rd} party only” insurance schemes, moving away from traditional “full cover” insurance policies due to economic pressure. The most vulnerable segment to this transition are customers who feel they have a lower risk of insurance claims. Insurance companies have a need 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.

Objective of machine learning model:

Purpose of the machine learning model is to classify existing and new customers based on the likelihood of applying for 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

DATASET

Car insurance dataset - Owner: Sagnik Roy - https://www.kaggle.com/datasets/sagnik1511/car-insurance-data?resource=download

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
CHILDREN	Boolean 1 or 0
POSTAL_CODE	Code corresponding to address location
ANNUAL_MILEAGE	Numerical value
VEHICLE_TYPE	Classified and Sedan or sports car
SPEEDING_VIOLATIONS	No of occurrences
DUIS	No of occurrences
PAST_ACCIDENTS	No of occurrences
OUTCOME	Boolean 1 or 0 based on request for insurance claim

METHODOLOGY

- ❑ Solution approach:
  - ✓ Binary classification model
  - ✓ Outcome of 1 if a claim is made and 0 if a claim is not made as the y variable
  - ✓ 17 features in data set that can be used

- ❑ Target models
  - ✓ Logistic regression model
  - ✓ Decision Tree Classifier
  - ✓ RandomForest Classifier
  - ✓ SVM model

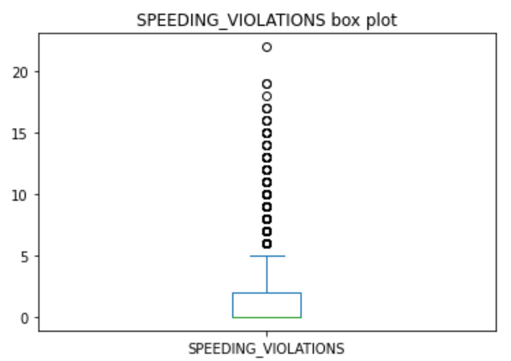
- ❑ Tools
  - Python
  - Google Colab
  - github

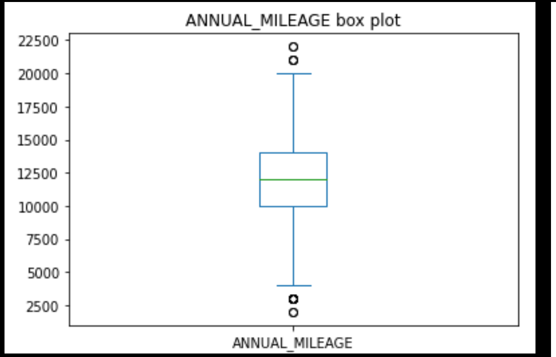
DATA PRE-PROCESSING

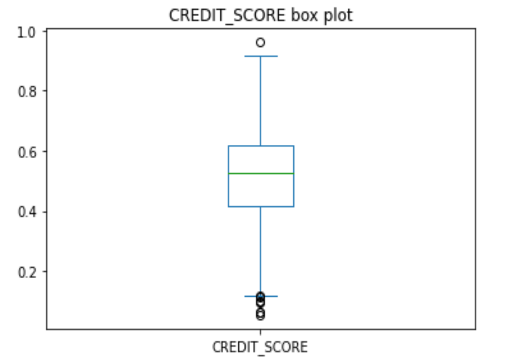
Drop rows with outliers

Drop rows with missing data

Encode categorical data using label encoding



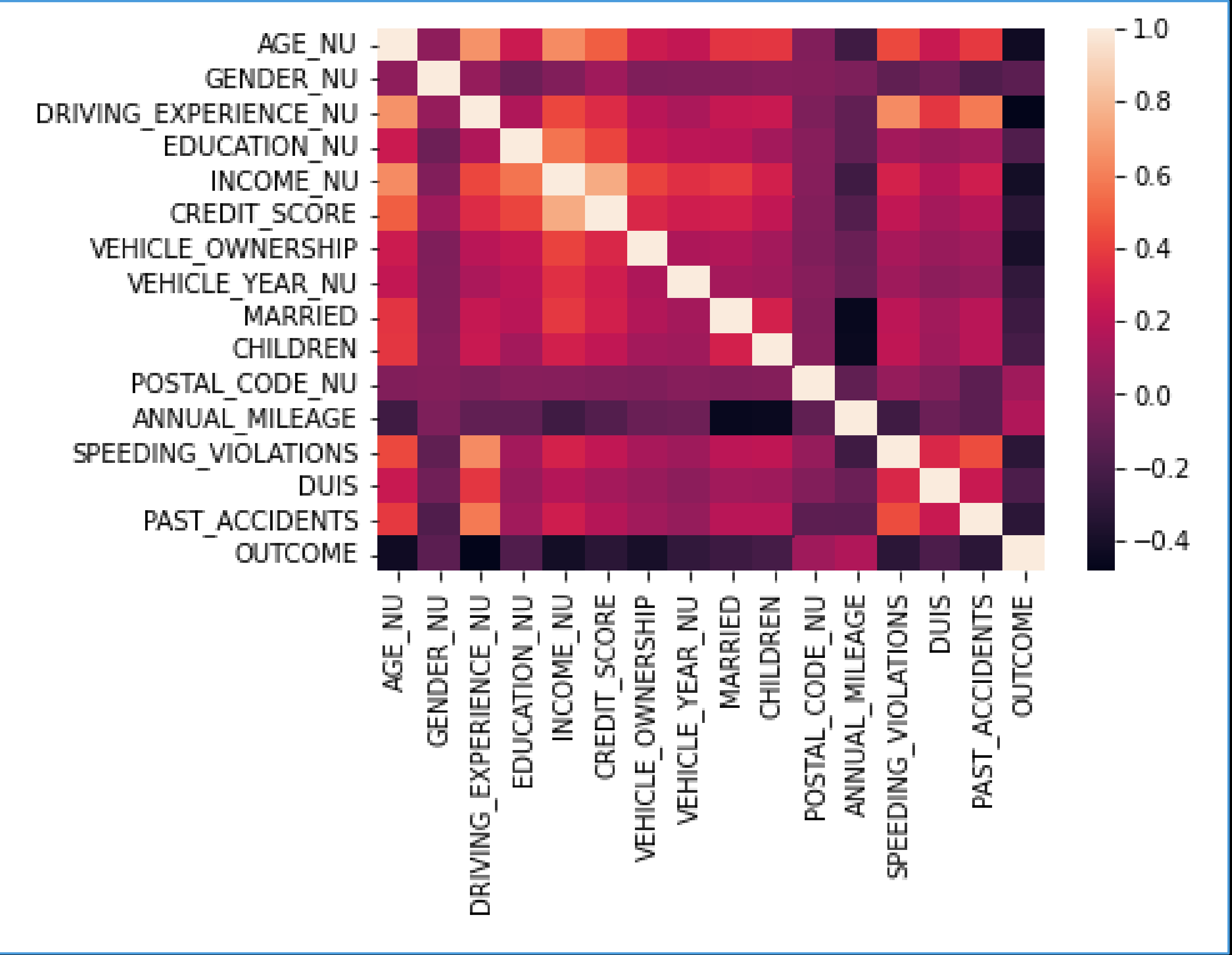




	AGE	AGE_NU	INCOME	INCOME_NU
7245	65+	4	upper class	4
507	26-39	2	poverty	1
7237	26-39	2	middle class	3
2006	16-25	1	poverty	1
9392	65+	4	upper class	4

MODEL TUNNING

- All the features selected show similar correlations to OUTCOME and are used in model tunning
- A train test split of 70:30 is used.
- The same training and test data are used for training and testing all 4 models
- Backward elimination is used to optimize the feature set



RESULTS

❑ Iteration A - Using 9 features

	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

Features:

- AGE_NU
- GENDER_NU
- DRIVING_EXPERIENCE_NU
- INCOME_NU
- CREDIT_SCORE
- VEHICLE_OWNERSHIP
- VEHICLE_YEAR_NU
- DUIS
- PAST_ACCIDENTS

❑ Iteration B – Using 13 features

	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

Features:

- AGE_NU
- GENDER_NU
- DRIVING_EXPERIENCE_NU
- INCOME_NU
- CREDIT_SCORE
- VEHICLE_OWNERSHIP
- VEHICLE_YEAR_NU
- DUIS
- PAST_ACCIDENTS
- MARRIED
- CHILDREN
- POSTAL_CODE_NU
- ANNUAL_MILEAGE

Iteration B with 13 features on RandomForest classifier provides the highest Accuracy, Precision and f1 score

CONCLUSION

- The final model has an accuracy of 84% and a precision of 81%

		Predicted	
		0	1
Actual	0	1331	130
	1	221	558

- The false negative rate is 14.2%, which ensures only 14 out of 100 people selected as low risk are likely to initiate insurance claims
- Therefore, the developed ML model can be used effectively to identify low risk users
- Based on feature importance driving experience has the highest importance of features

	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