Analysis & Prediction of Car Insurance Claims

Nelette Louw

Table of Content

Stakeholder and Business Problem

- **The stakeholder :** An insurance company

- **Business Problem**: The target is to analyse and predict customer behaviour i.e. whether the customer will claim against their car insurance or not.

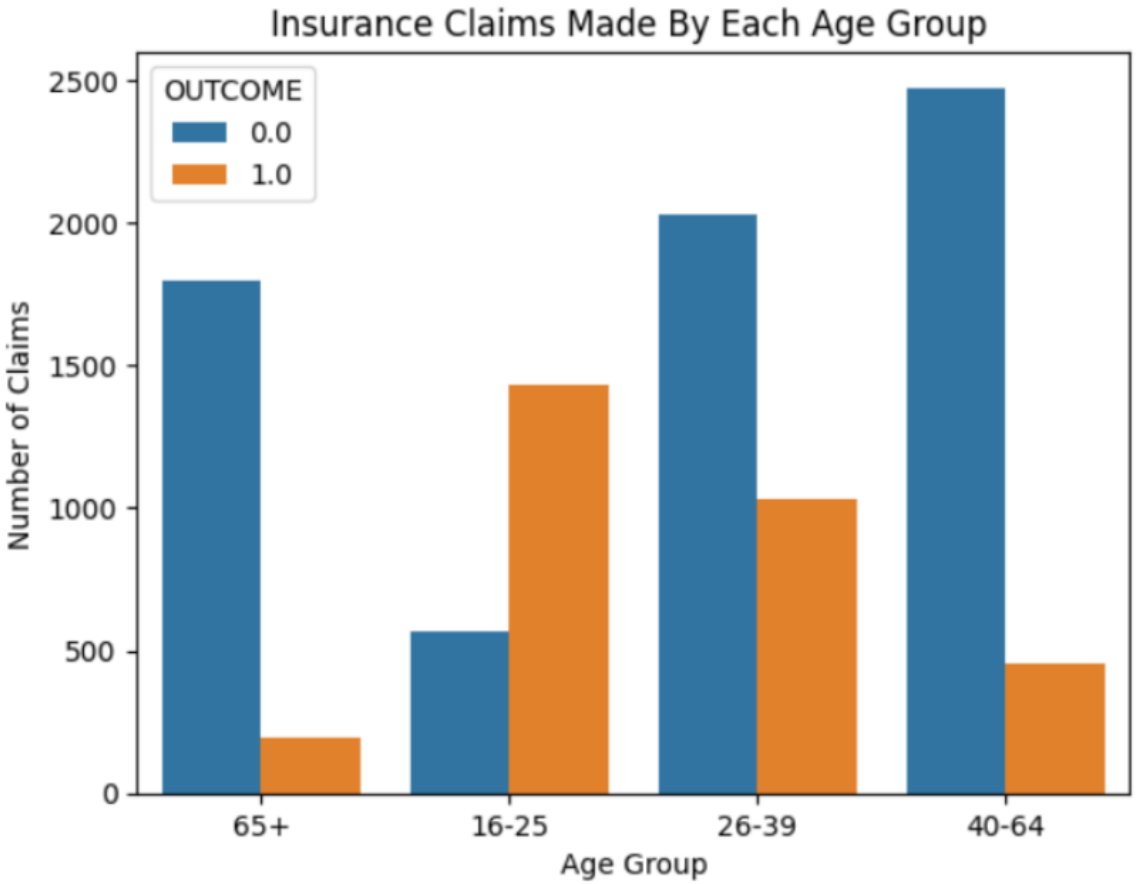
Information about the Data

- The annual car insurance data consists of 19 columns and 10 000 rows.
- Each row represents the data of a customer and the column **"Outcome"** indicates whether the customer has made an insurance claim or not (1 = Claimed, 0 = Not Claimed)

ID	AGE	GENDER	RACE	DRIVING EXPERIENCE	EDUCATION	INCOME	CREDIT SCORE	VEHICLE OWNERSHIP	VEHICLE_YEAR	MARRIED	CHILDREN	POSTAL CODE	ANNUAL MILEAGE	VEHICLE TYPE	SPEEDING VIOLATIONS	DUIS	PAST ACCIDENTS	OUTCOME
569520	65+	female	majority	0-9y	high school	upper class	0.629027	1	after 2015	0	1	10238	12000	sedan	0	0	0	0
750365	16-25	male	majority	0-9y	none	poverty	0.357757	0	before 2015	0	0	10238	16000	sedan	0	0	0	1
199901	16-25	female	majority	0-9y	high school	working class	0.493146	1	before 2015	0	0	10238	11000	sedan	0	0	0	0
478866	16-25	male	majority	0-9y	university	working class	0.206013	1	before 2015	0	1	32765	11000	sedan	0	0	0	0
731664	26-39	male	majority	10-19y	none	working class	0.388366	1	before 2015	0	0	32765	12000	sedan	2	0	1	1

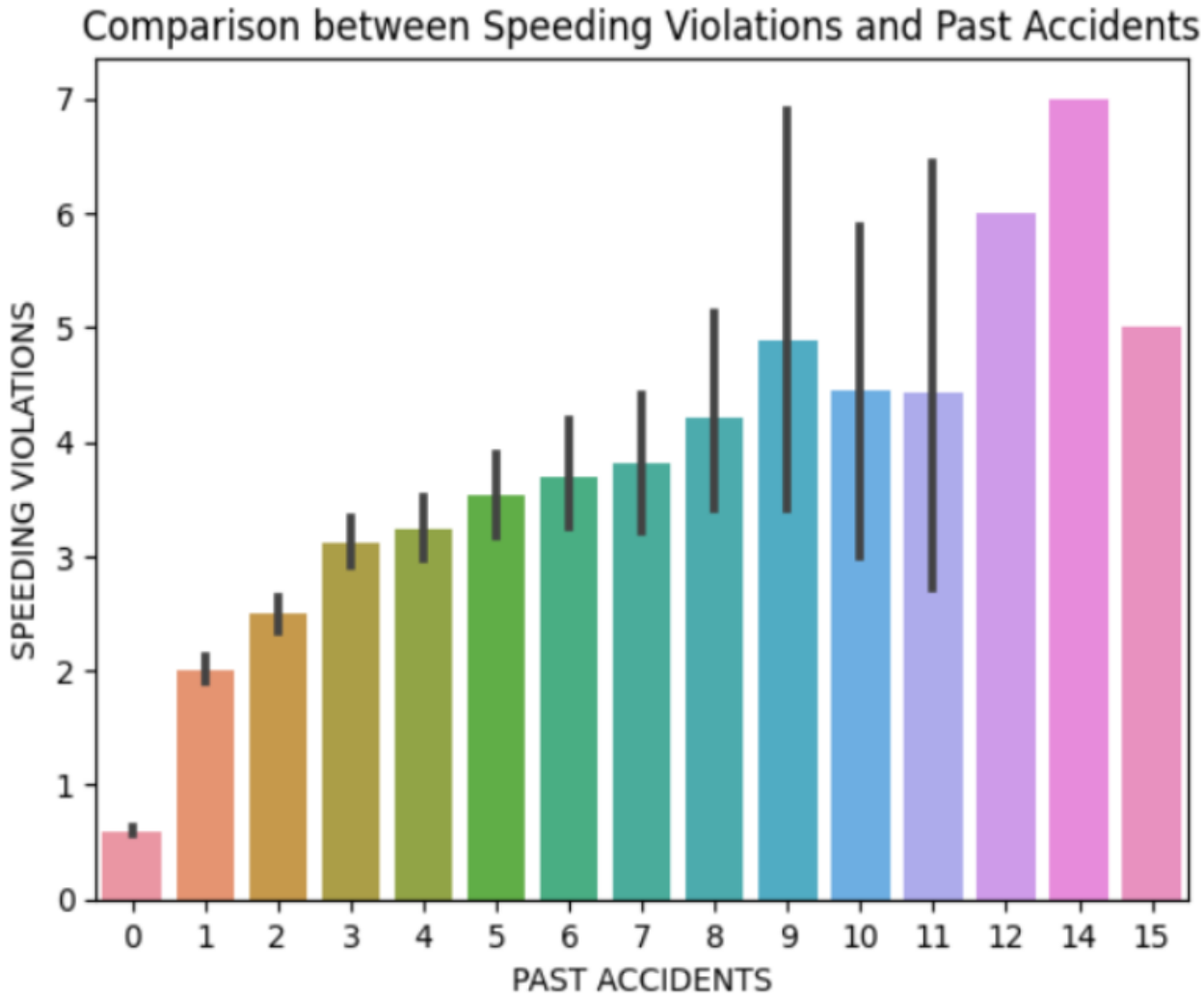
Approach / Method

Visualization 1



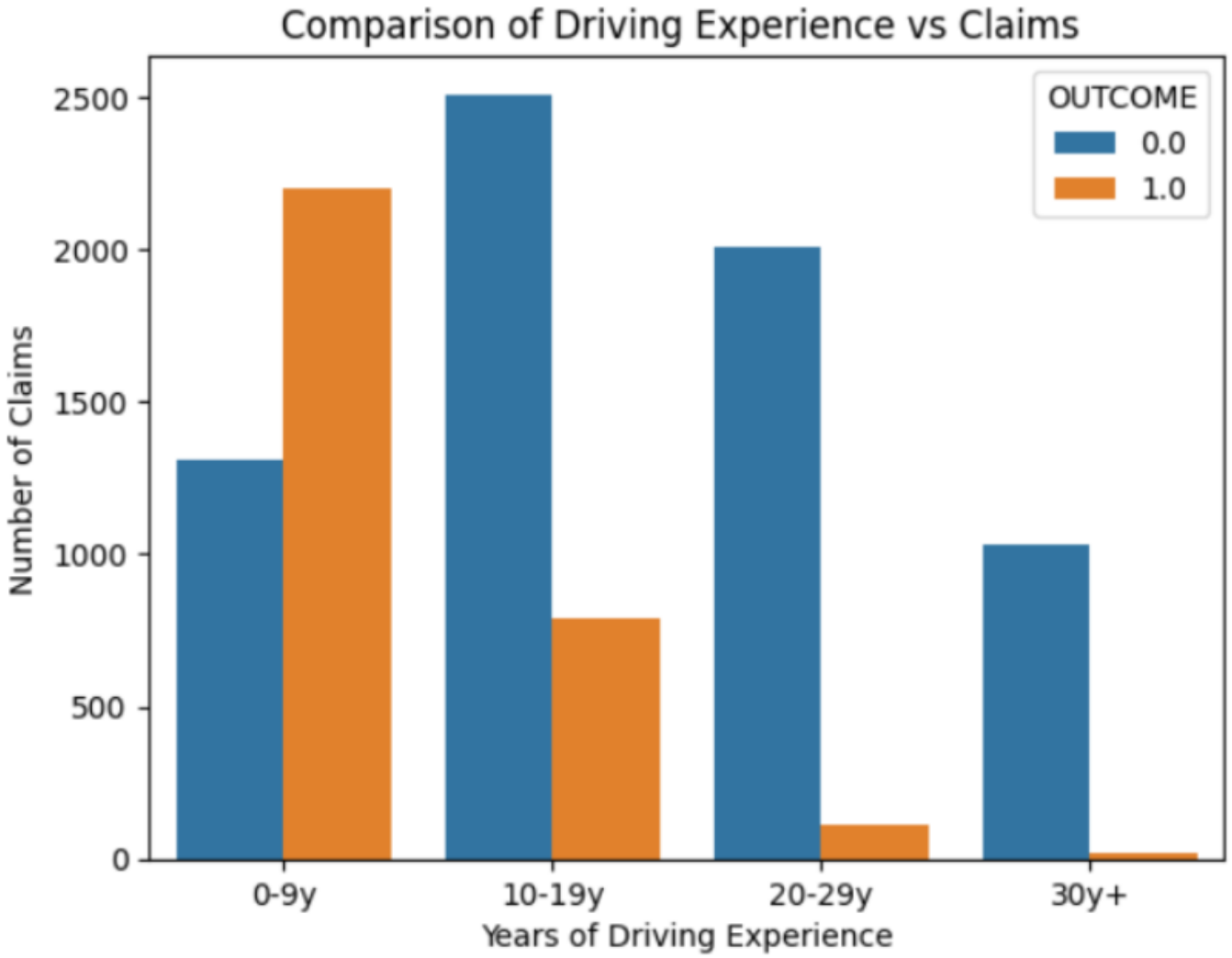
More `Insurance claims` are made by the `Age Group` 16-25 year than other age groups.

Visualization 2



When Past Accidents and Speeding Violations are compared, it shows that the more Speeding Violations customers have, the more likely it is that they also have been involved in Past Accidents.

Visualization 3



Claims made by customers gets lower with the more `Years of Driving Experience` a customer has. Customers who fall in the bracket `10-19 years of Driving Experience` shows to make the least claims overall.

Strengths and Limitations

Machine Learning = is the process of automatically learning from data without requiring explicit programming, with the ability to expand the knowledge learned with experience.

Strengths

- 3 different machine learning models were used to train and test data (KNN, Logistic Regression and Logistic Regression with PCA) to get the best results.
- The model Logistic Regression with PCA delivered the best results with an accuracy of 86%.

Limitations

- There might be more data elements/factors needed to accurately predict car claims than what is currently available in the insurance companies' data set.

Recommendations

- It is recommended that the insurance company should pay close attention to all the correlations and relationships that's been shown in the visualizations when determining the insurance premium of a customer.

- Customers in the age group 40-64 year will also mostly be the customers with the most years of driving experience. To enhance and attrack more business, this age group's premiums could potentially be discounted.

Nelette Louw

Nelette.Louw@mmltd.co.za

Thank you