Predicting Fraudulent Claims in Automotive Insurance By Leveraging Machine Learning

Presented by

Manu M L

Trainee Data Scientist at Turing Minds.AI

Content

1. Case Introduction
2. ML Problem Statement
3. Data Exploration and Visualization
4. Data Pre-processing
5. Important Features
6. Models
7. Validation and Parameter Tuning
8. Conclusion

Case Introduction

- The global insurance market grew at a rate of 450 billion dollars from 2021 to 2022
- Types Of available insurances: Health Insurance, Motor Insurance, Home Insurance, Fire Insurance
- Major Challenges faced by the sector: Customer retention, Rising operation costs, Fraudulent claims and many more.
- The companies have lost $6.25 billion due to Fraudulent claims.

ML Problem Statement

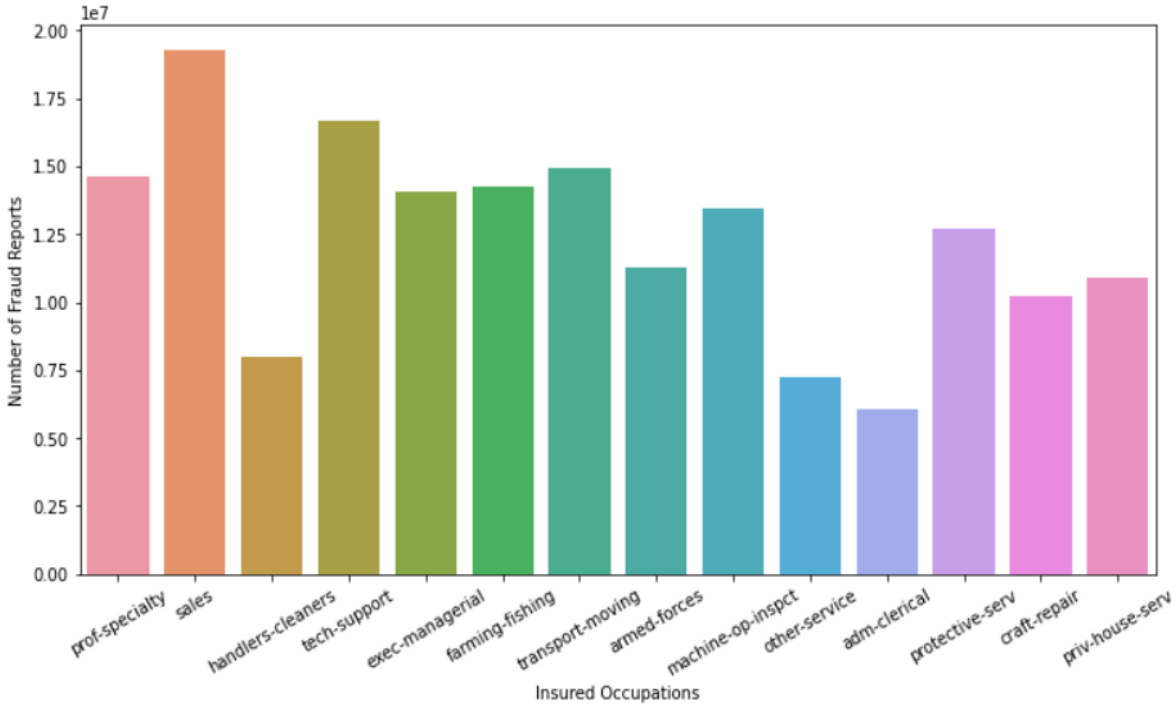
- The problem
- Why does it need to be solved?
- Why ML?

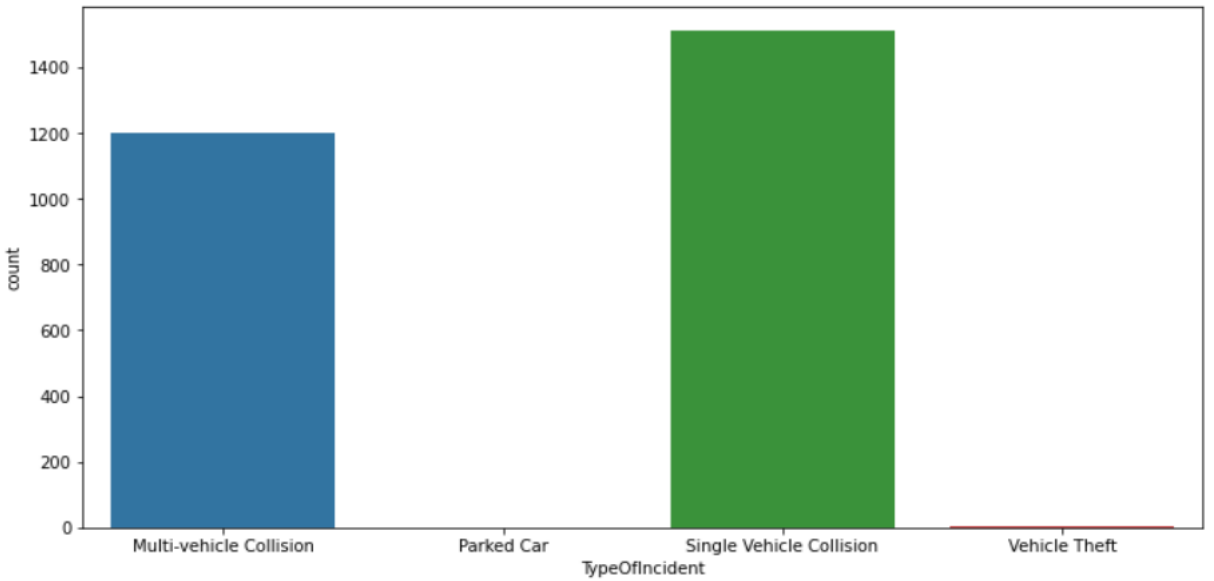
Data Exploration and Visualization

Given datasets: Claim data, Customer details, Policy data, Vehicle details and Reported Frauds.

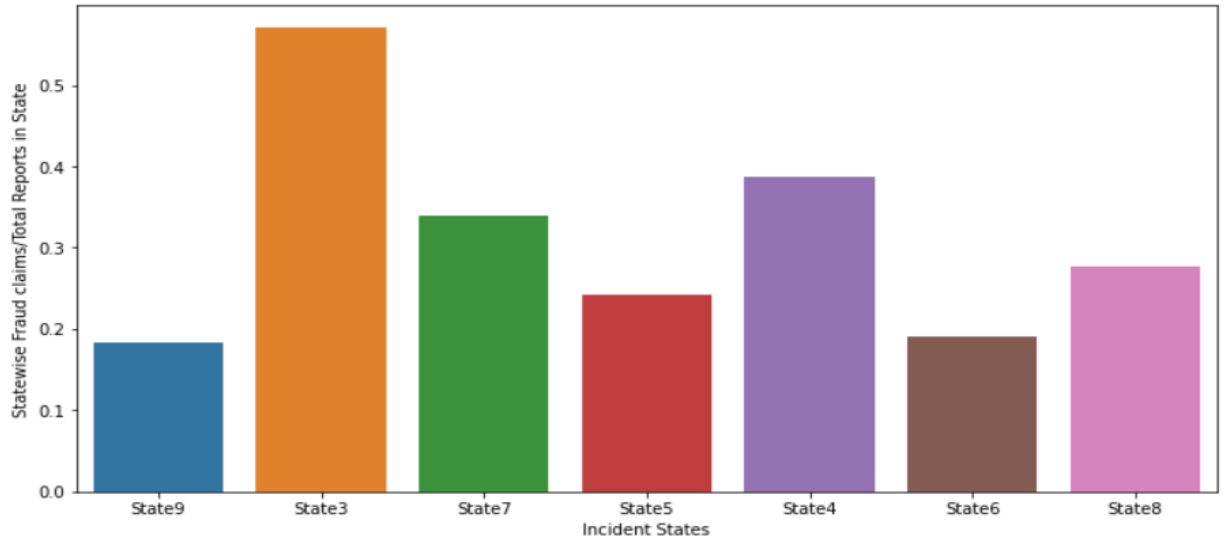
Using Visualization tools to:

- Understand data distribution
- Obtain insights and traits
- Identify features that are important that might have major influences

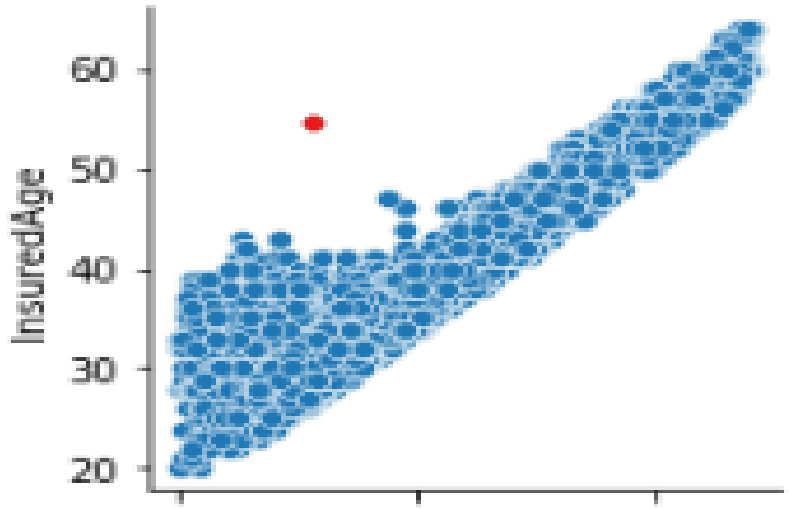




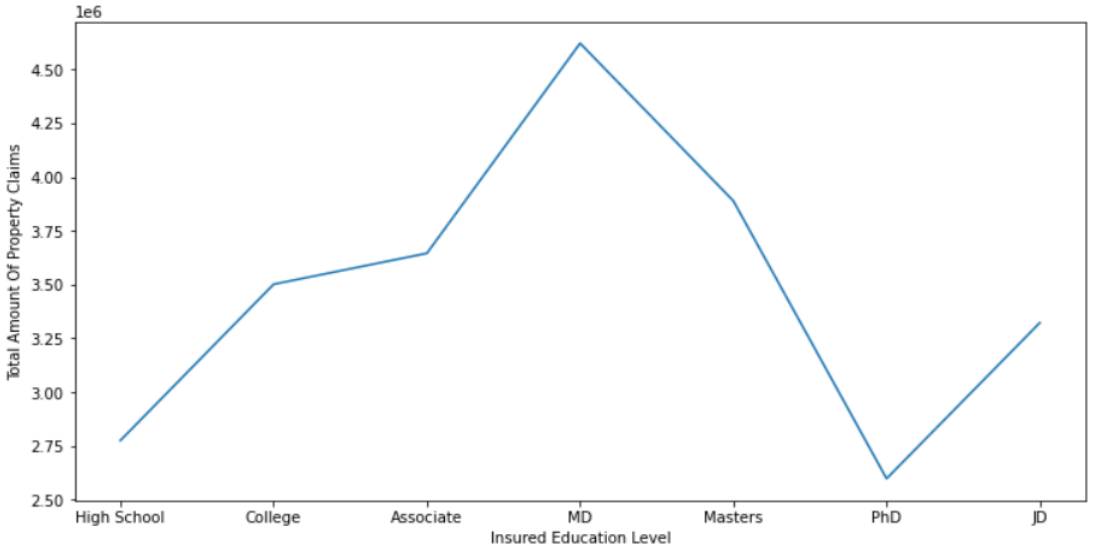

```
The Amount of total fraud claim in Single Vehicle Collision 98881243.0
The Amount of total fraud claim in Multi-vehicle Collision 74569402.0
The Amount of total fraud claim in Vehicle Theft 39736.0
The Amount of total fraud claim in Parked Car 53044.0
```

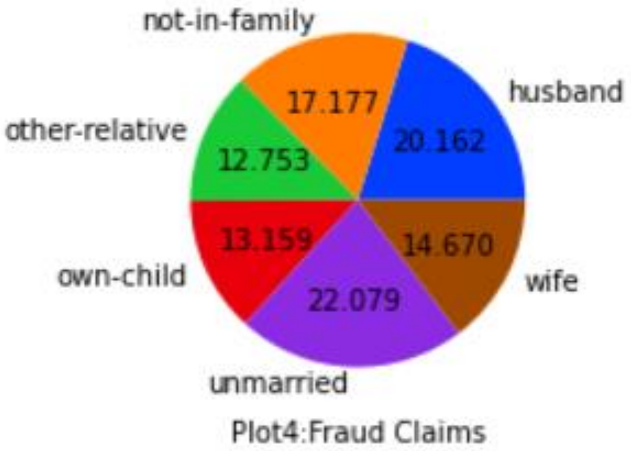


Plot6:Statewise ratio of Fraudulent claims

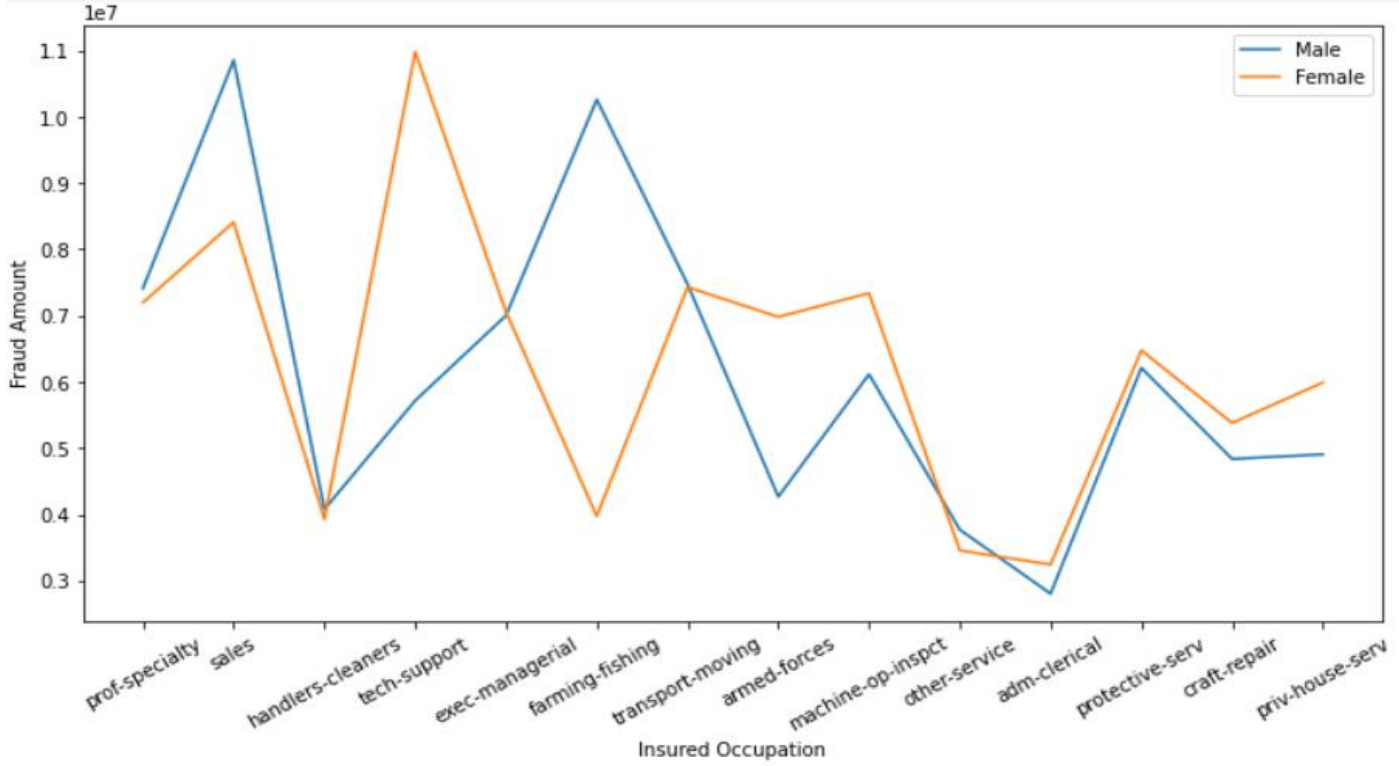


Customer Loyalty Period





Plot4:Fraud Claims



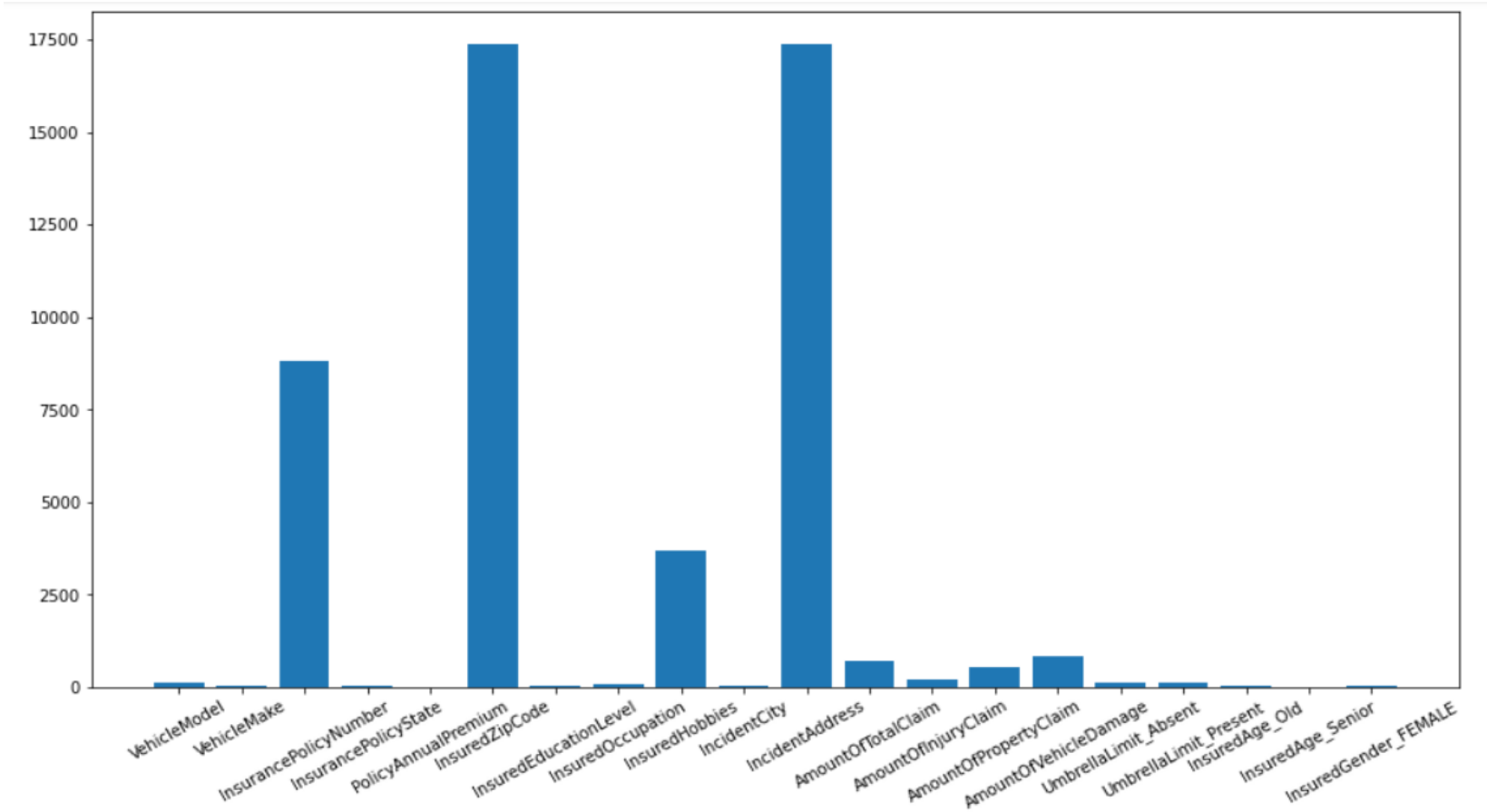
Plot9

The Sum of Fraud claim amount by Male: 85706801.0
The Sum of Fraud claim amount by Female: 87836624.0

Data Pre-processing

Steps	Methods
Missing Value Treatment	Simple Imputer: mode Iterative Imputer
Outlier Detection	Flooring Multiplication with -1
Feature Engineering	Delta Days
Feature Transformation	Numerical to Categorical: Binning Categorical to Numerical: Response Encoding, One Hot Encoding
Feature Scaling	Robust Scaler
Feature Selection	ANOVA F-value

Important Features



Models

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Light Gradient Boost Classifier
5. Voting Classifier : KNN, Decision Tree, Logistic Regression, Gaussian Naïve Bayes, SVC

Validation and Parameter Tuning

```
estimators=[LogisticRegression(max_iter=500),DecisionTreeClassifier(),RandomForestClassifier(),
            lgbm.LGBMClassifier(),VotingClassifier(estimators=estimators_vote,voting='hard')]
rsk=RepeatedStratifiedKFold(n_splits=5,n_repeats=5)
score=cross_val_score(estimator=i,X=X,y=y,scoring='f1_weighted',cv=rsk,verbose=False)
```

```
RandomForestClassifier(max_depth=12, n_estimators=500)
KNeighborsClassifier(n_neighbors=7, p=1)
LogisticRegression(C=10, class_weight='balanced')
DecisionTreeClassifier(max_depth=4, min_samples_split=4)
LGBMClassifier(learning_rate=0.001, max_depth=12, n_estimators=50
0)
 SVC(C=10)
```

Conclusion

Model Summary				
Estimator	**F1 score(Validation)**	**Run Time**	**ROC_AUC score**	**Test Score**
Logistic Regression	0.9380	00.3475s	0.9060	0.68
Decision Tree Classifier	0.8839	00.1397s	0.9329	0.7
Random Forest Classifier	0.9413	14.1836s	0.9305	0.84
LGBM Classifier	0.9422	10.0978s	0.9327	0.82
Voting Classifier	0.9418	259.5843s	0.9289	0.83

The model was able to predict fraudulent claims with **94% accuracy** on the train data and F1 score of **0.84 on unseen test data**

Genuine claims = 2065/2105 (predicted/actual)(train data)

Fraudulent claims = 655/779 (predicted/actual)(train data)

Conclusion : Recommendations to Business

- To have thorough verification of background for the claims in the name of “Husband” and “unmarried”.
- Higher the age of the customer more is the loyalty period and hence are to be given more importance compared to younger ones.
- To re-evaluate the amount that can be claimed under Minor damage as there are claims that are worth more than those from Major damage and Total Loss.
- To examine the claims thoroughly from customers working in Sales, Transport-moving and tech-service.
- To have a thorough check on claims when the Incident State is State3 followed by State4.

Conclusion

What more could have been done?

A further detailed study and discussion with the domain expert regarding the given features might have helped to understand the dataset better and also would have been useful for feature engineering.

Thank You