“Car Insurance Claim Prediction”

By Ashish Nautiyal

Problem Statement

Develop a predictive model that assesses the claim probability for car insurance policies.
The objective would be to understand the factors that influence claim frequency and enable insurance companies to better assess risk and determine appropriate premiums for policyholders.

About Data

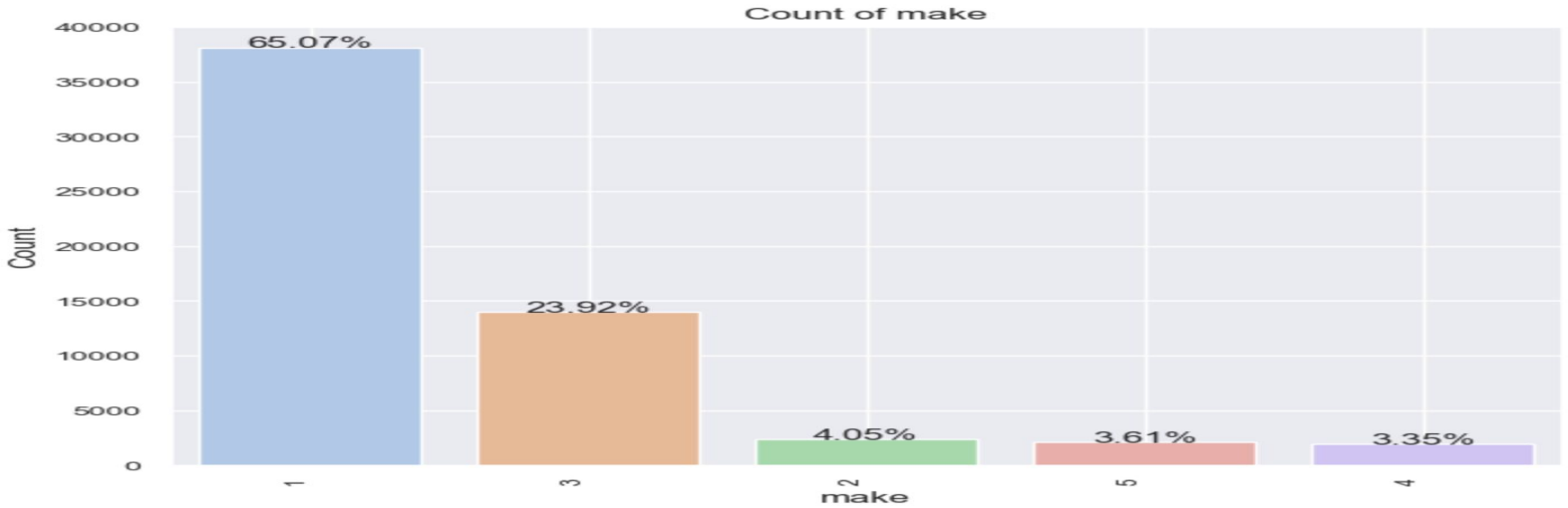
Features	44
Total no of records	58592
Type of data	Float, Integer, String
Nan of records	0
Duplicate Records	0
Target Feature(y)	is_claim

Data Cleaning and Preprocessing

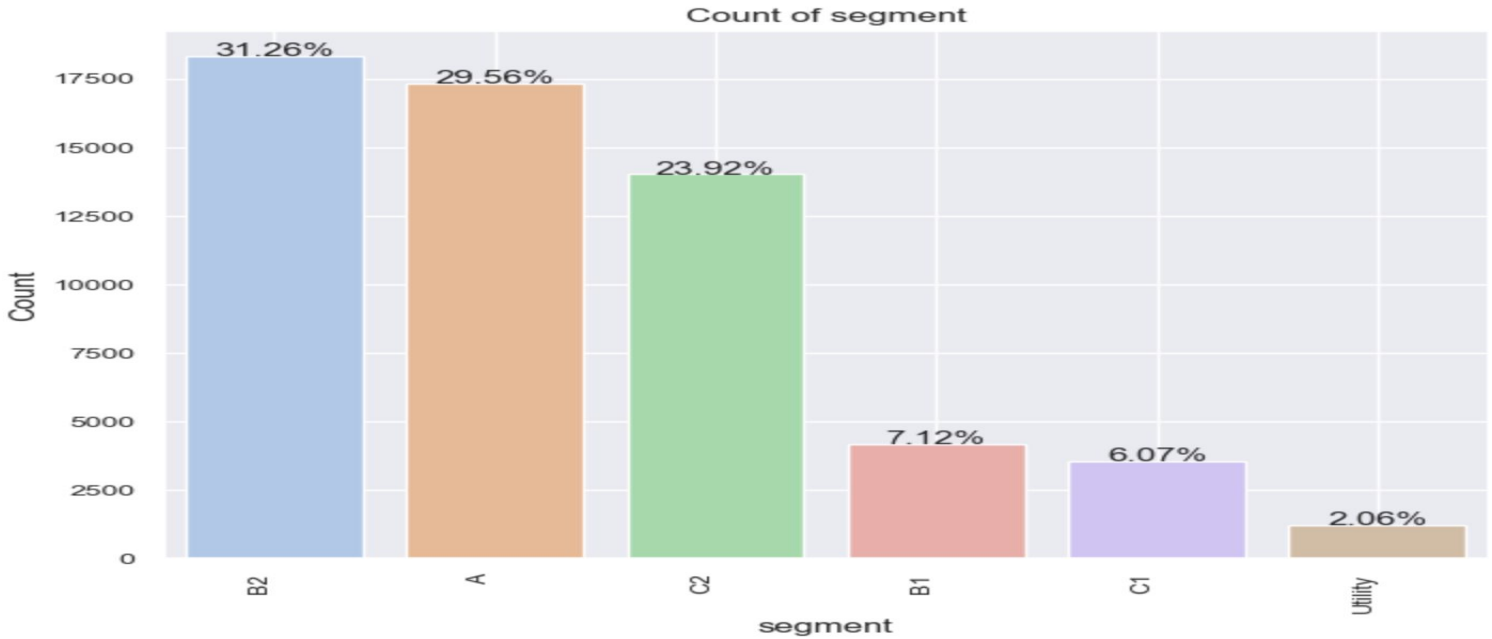
- No Missing values in data.
- Data type conversion was done as per feature compatibility.
- For outliers treatment we used simple capping method.
- Encoding label encoder is used.
- Data Imbalance is treated by SMOTE.

Insights

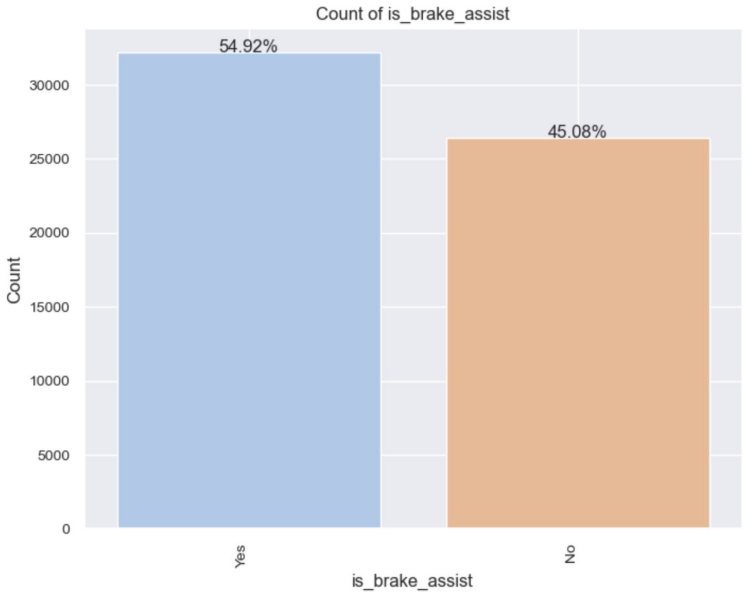
- Most of the vehicles in market are of type 1.Make 3 is on 2nd number.

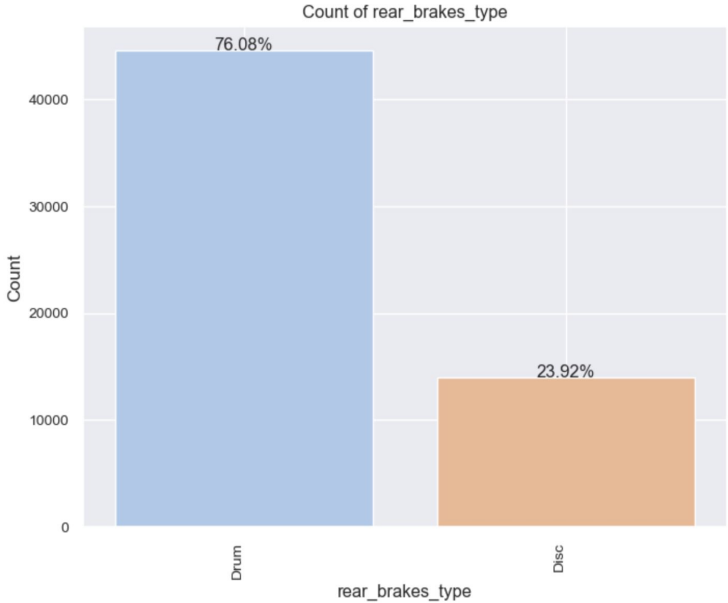


- B2 is no 1 Segment which has 31% ,A and C2 are other segment leaders.
- Utility vehicles are less in number.

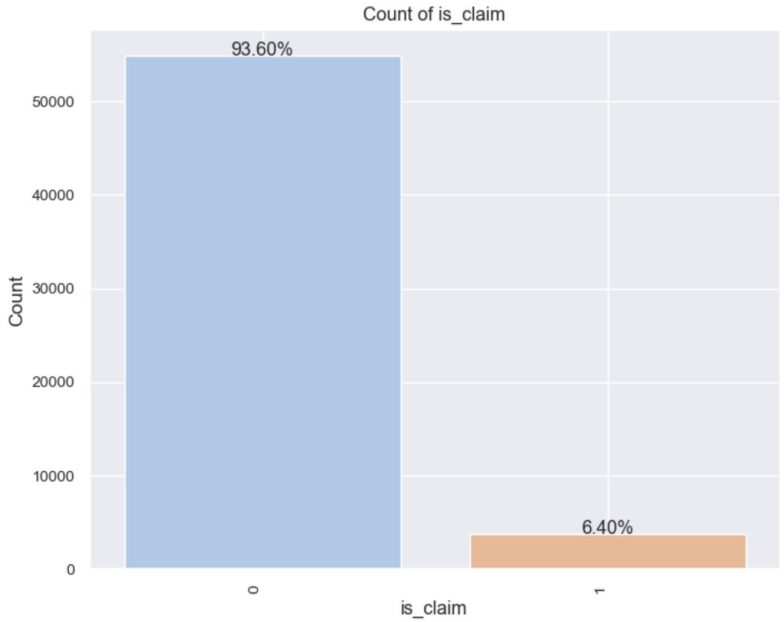


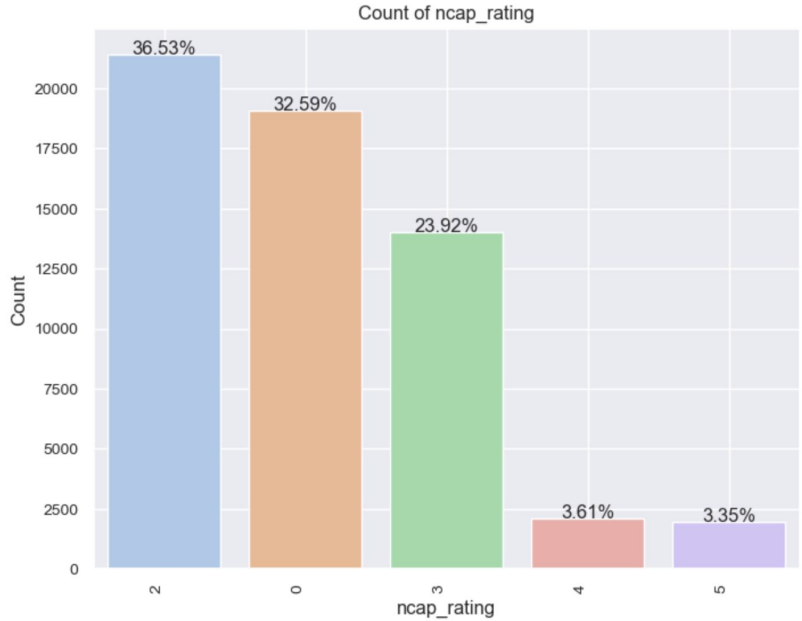
- Brake assist feature is available in 55% vehicles.
- 77% vehicles have drum brake which are not so efficient in case of emergency.



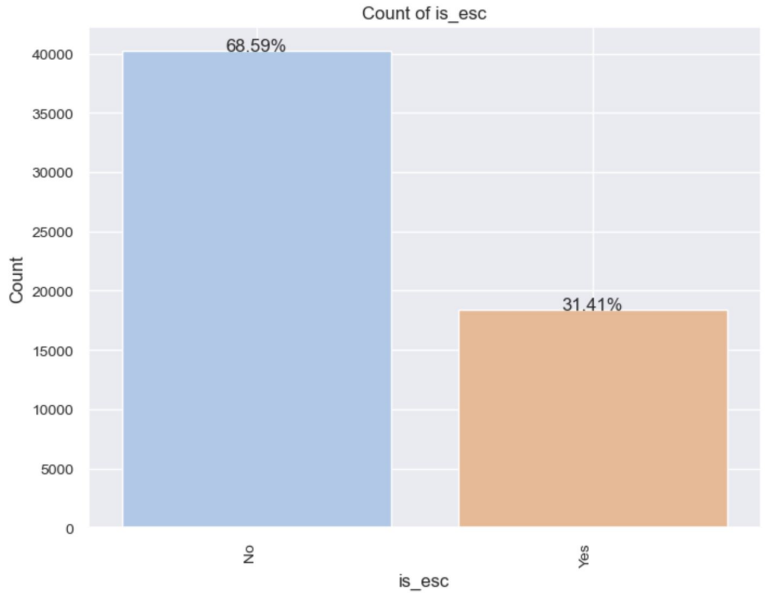


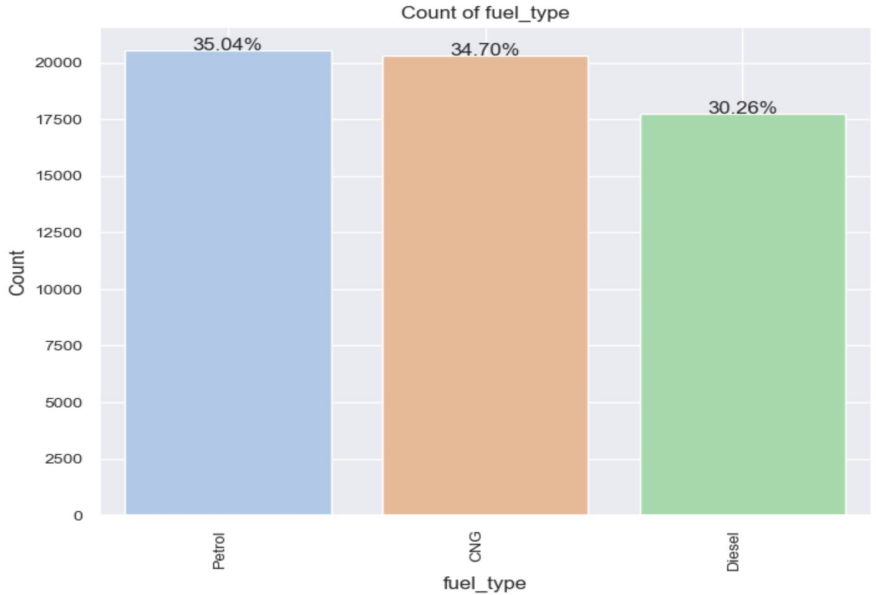
- 93% times insurance was not claimed and only 6.4% is claim.Out of 100 6.5 claims are raised for insurance.
- 74% of vehicles have less than 3 Ncap rating and only 7% have 4,5 Ncap.



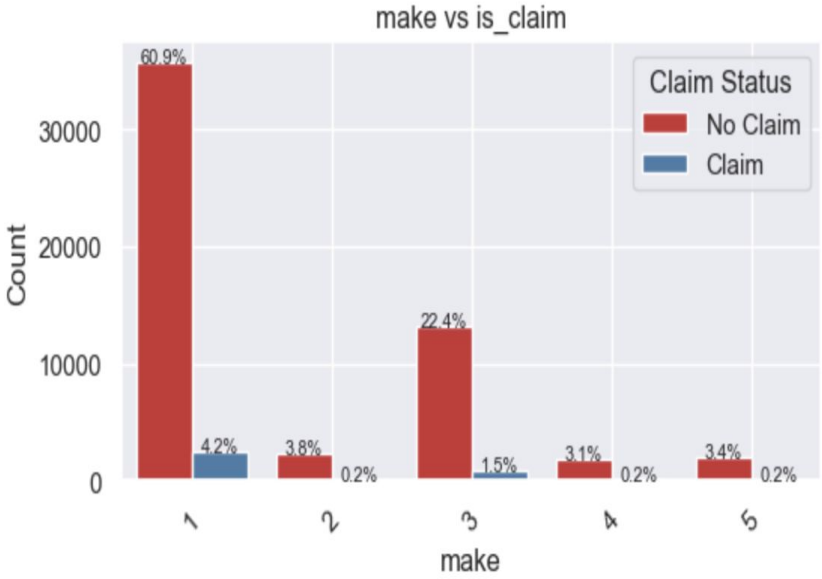


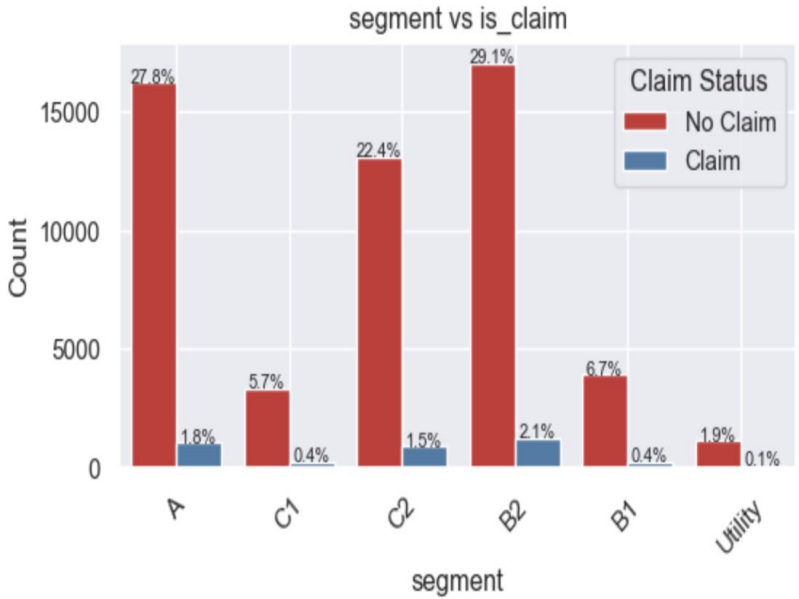
- 69% of vehicles don't have ESC which is a critical safety feature.
- Fuel type wise petrol and cng are kind of similar as 35 and 34 % and 30% diesel vehicle.



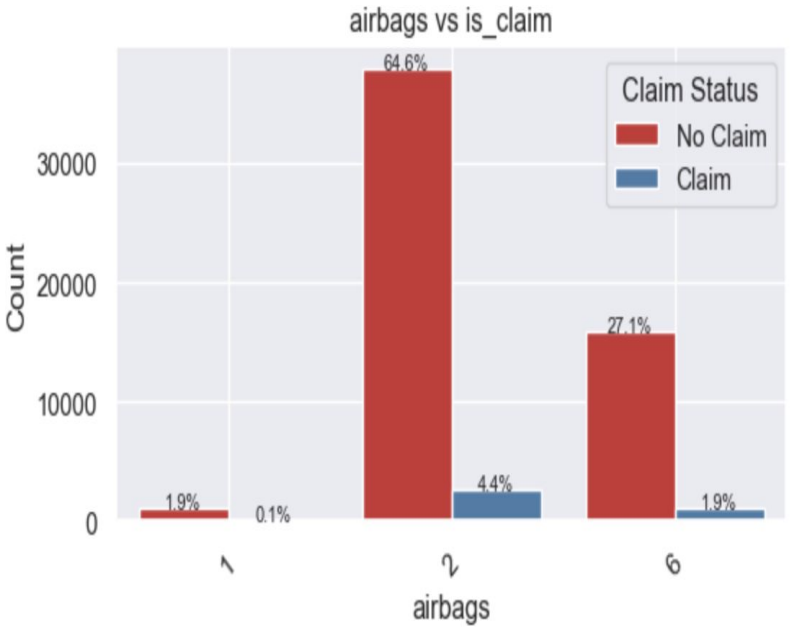


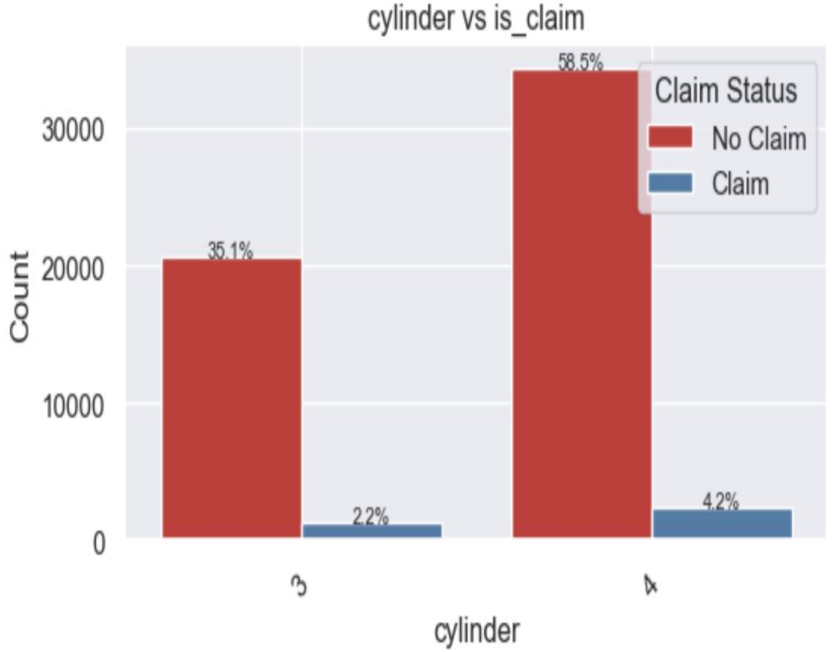
- Make 3 and 1 vehicles have major chunk of claims.
- Segment wise A,B2,C2 have more claims.This depends on total number of market share also.





- 6 airbag vehicles have 2% claim and 2 airbag vehicles have 4.4% claims
- 4 Cylinder vehicles have more % of claims as compared to 3 cylinder





Feature Importance

From Statistical test like Chi2 ,Correlation,Anova and Random forest model feature importance,We analysed that below are top 5 most significant features for claim prediction

	feature	rf_importance
0	policy_tenure	0.475934
2	age_of_policyholder	0.263380
1	age_of_car	0.153861
4	population_density	0.050274
3	area_cluster	0.048483

Model Building

In this Classification problem we have tried 3 ensemble models and below are the result we got:

Model Name	Training Set Accuracy	Testing Set Accuracy
Random Forest	99.9%	93.7%
Gradient Boost	94.2%	93.5%
XgBoost	94.3%	92.7%

- With 10 fold cross validation we are getting mean accuracy around 93%.
- Here i am selecting Gradient boost as it has less fluctuations and seems stable by results.

Further Improvements/Suggestions:

- We can go with limited features as per feature importance and check whether there is an improvement in model.
- We can also integrated Pipelines which i have not included in this iteration for preprocessing and model building.

THANK YOU