Predicting Car Insurance Claims

Orion Joyner
CSC 461 Machine Learning

Introduction

Goals:

- Are policy claims predictable within the next six months?
- What factors specifically contribute to the policy holder?

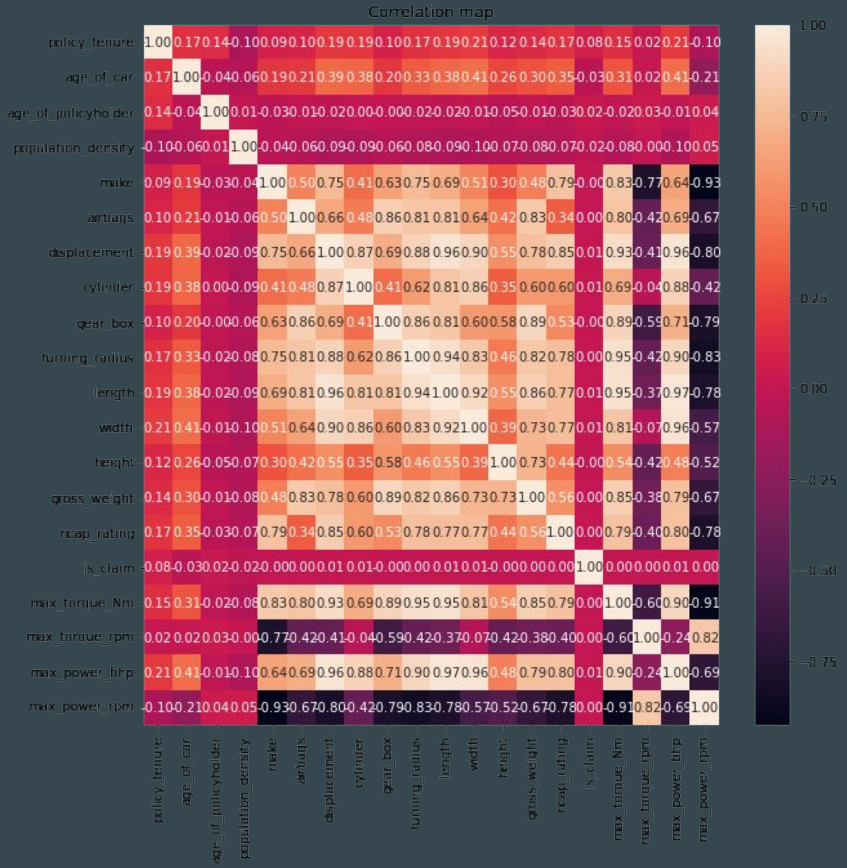
Introduction

Approaches:

- Some of the approaches that we used for our project included cleaning the initial data to drop missing values that don't correlate with the wanted result.
- Training and visualizing the data that can be seen appropriately.
- Using methods such as Confusion Matrices, Decision Tree Classifiers, GridSearch and resampling our data for precision and Accuracy.

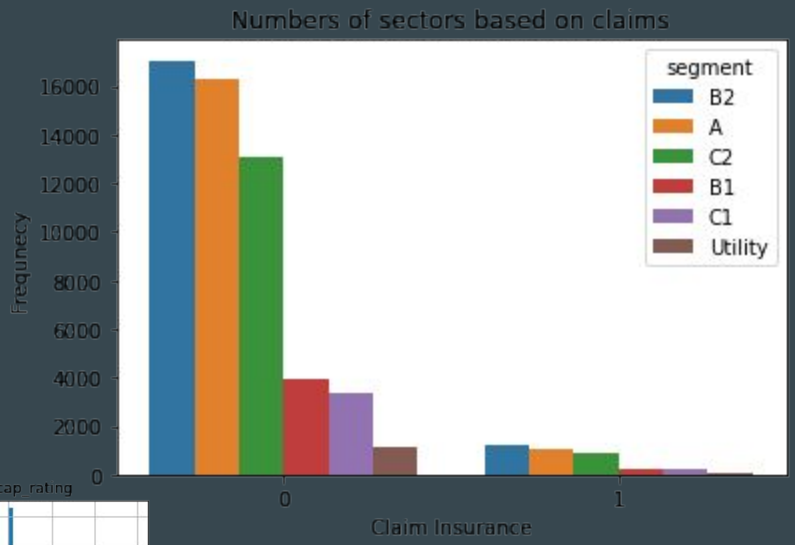
EDA(Data Analysis)

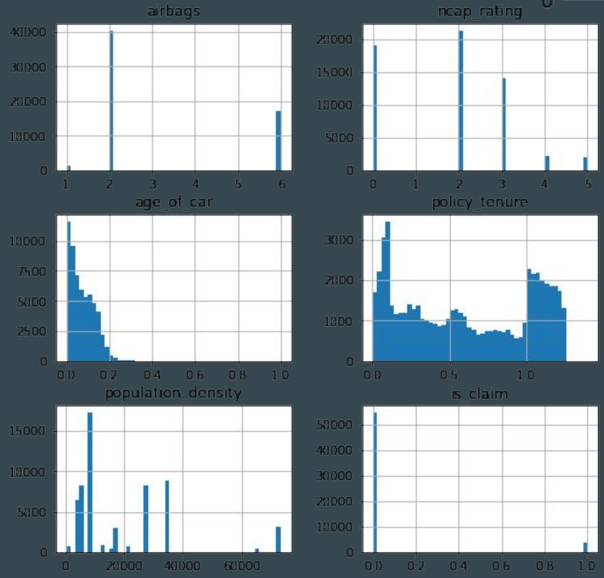
- After gathering the data, the goal was to identify our features and what relate to one another. Our approach for this was to construct a correlation matrix/map to identify these features.
- By referring to the figure to the right, most of the features correlate to each other, mainly policy_tenure, age_car, age_policyholder, population density, and is_claim

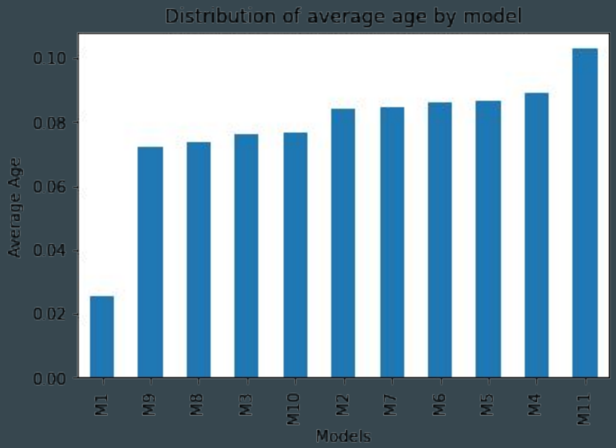


EDA(Continued

- Continuing our EDA, we were able to determine what main factors contribute the most to the policy_holders claim.
- The the right, you'll see the different approaches we took to for our EDA.
- The six major contributors were airbags ncap_rating age_of_car policy_tenure population_density is_claim







Base Model and Model Tests

Basemodel:

Accuracy score = 0.4964423489069867
F1 score = 0.11268882175226586
Recall score = 0.49866310160427807
Precision score = 0.06352179836512262

Final model:

Accuracy score = 0.5873124732104587
F1 score = 0.16481609993060375
Recall score = 0.6350267379679144
Precision score = 0.0946969696969697

```
Accuracy score =
0.4964423489069867

AUC-ROC score =
0.4974766456084045

F1 score =
0.11268882175226586

Recall score =
0.49866310160427807

Precision score =
0.06352179836512262
```

Conclusion

- The final model is quite inaccurate. Of course in comparison with base model final is better a little bit. I tried a lot of kinds of deal with multicollinearity(just do nothing, PCA), sampling(Randoms, class balance). However in the end I chose the best one to show.