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.

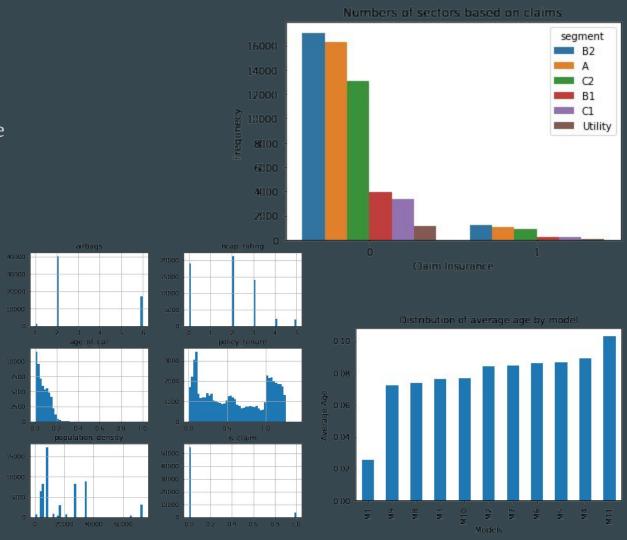
EDA(Data Analysis)

- 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

```
1 00 0 170 140 100 090 100 19 0 190 100 17 0 190 210 12 0 140 170 080 15 0 020 21-0 10
age of car -0.17<mark>1.00-</mark>0.040.060.190.210.390.380.200.330.380.410.260.300.350.030.310.020.41-0.21
            0.09 0.190,030,041,00 0.500,75 0.410,630,75 0.69 0.510,30 0.48 0.79 0.00 0.83 0.77 0.64 0.93
            0.10 0.21-0.010.060.50 1.00 0.66 0.48 0.860.81 0.81 0.64 0.42 0.83 0.34 0.00 0.80 0.42 0.69 0.67
  cylinder -0.19 0.38 0.00-0.09 0.41 0.48 0.87 1.00 0.41 0.62 0.81 0.86 0.35 0.60 0.60 0.01 0.69 0.04 0.88 0.42
            0.10 0.20-0.000,060.630.860.69 0.411.00 0.86 0.810.600.58 0.89 0.5
            0.17 0.33 0.020 08 0.75 0.81 0.88 0.62 0.86 1.00 0.94 0.83 0.46 0.82 0.78 0.00 0.95 0.42 0.90 0.83
            0.19 0.38-0.020.09 0.69 0.81 0.96 0.81 0.81 0.94 1.00 0.92 0
             0.21 0.41 0.010,100.51 0.64 0.90 0.86 0.60 0.83 0.921 00 0.39 0.73 0.77 0.01 0.81 0.07 0.96 <mark>0.57</mark>
            0.12 0.26-0.050.070.300.420.55 0.350.580.46 0.550.39 1.00 0.73 0.44-0.000.5
              14 0 30 0 010 08 0 48 0 83 0 78 0 60 0 89 0 82 0 86 0 73 0 73 1 00 0
             ) 17 0.35-0.030.07<mark>0.79</mark> 0.34<mark>0.85</mark> 0.60 0.53 0.78 0.77 0.77 0.44 0.56<mark>1.00</mark> 0.00<mark>0.79-</mark>0.400.80<mark>-0.78</mark>
             0.15 0.31-0.020.08 0.83 0.80 0.93 0.69 0.89 0.95 0.95 0.81 0
            0.02 0.02 0.03 0.00 0.770 420 41 0.040 590 420 370 070 420 380 400 00 0 0.60
             0.21 0.41-0.010.10 0.640.69 0.96 0.88 0.71 0.90 0.97 0.96 0
            0.100.210.040.05-0.930.670.800.420.790.830.780.570.520.670.78
```

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.