Classifying Policyholders with Supervised Learning

By Ben Cox

Problem Statement

- The problem statement given to by my manager is the following:
- 1. Target the potential customers in 2022 which will result in the lowest claim amounts. That is,
- Predict the number of claims for each potential customer, and
- Predict the cost per claim, given that a claim occurs

The Data

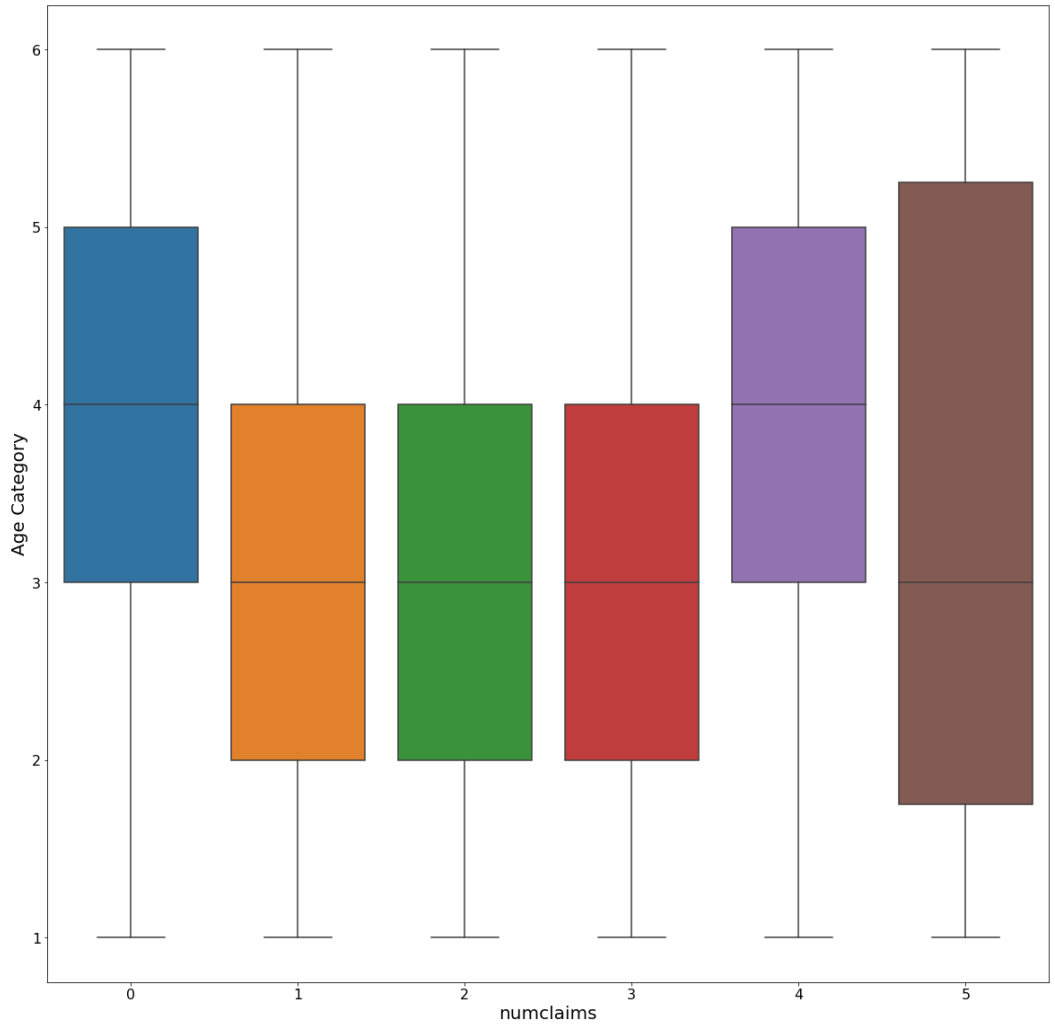
- I have received two datasets to analyze: a training dataset and a testing dataset.
- The training data set is a set of personal auto insurance policies taken out in 2020. There are 60,392 policies (rows), of which 10,030 had at least one claim. There is also a substantial number of missing values.
- The testing data set is a set of personal auto insurance quotes for insurance premiums starting in 2022. There are 7,464 quotes (rows). There is also a substantial number of missing values.

Potential Customers

- Predictive modeling was done to classify customers based on the number of claims and the cost of claims.
- Customers ranged from having 0 to 6 claims, with most of them being in the 0-1 claims range.

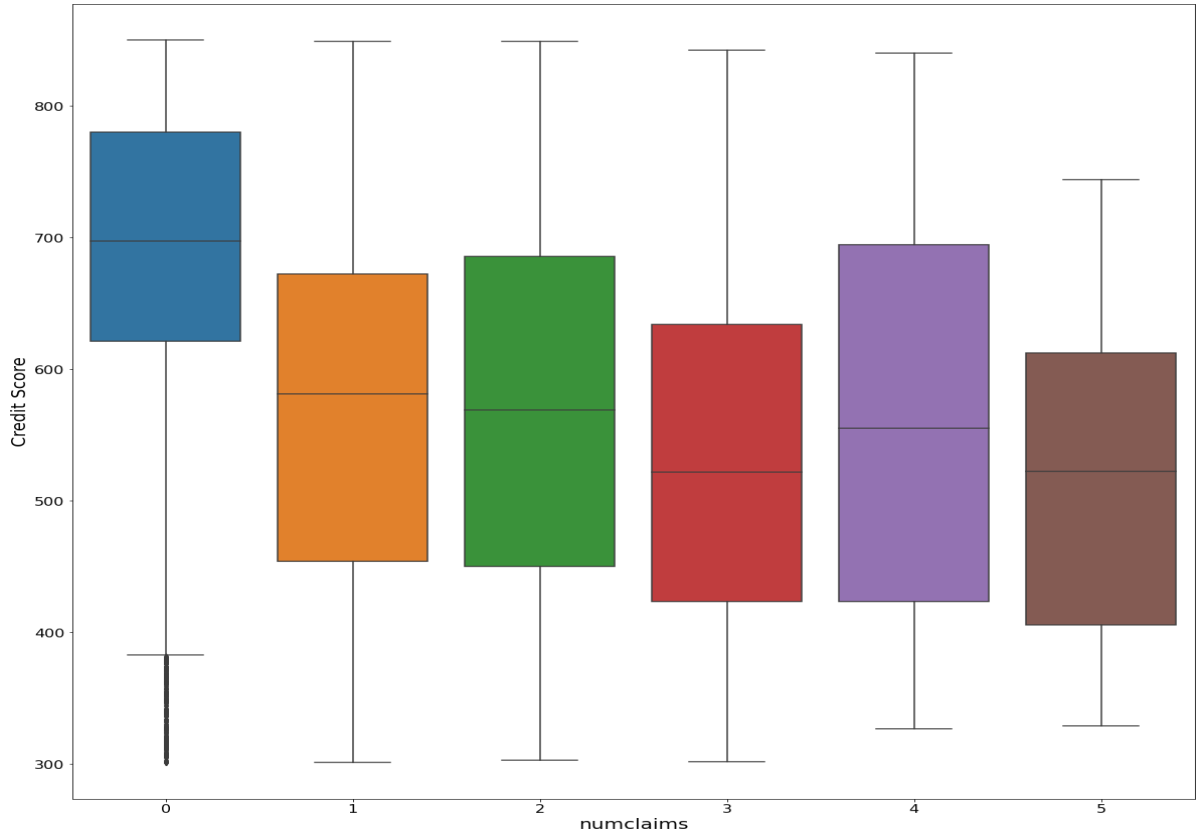
Age Category

- The drivers with no predicted claims had the highest median age besides the group of drivers with four predicted claims.



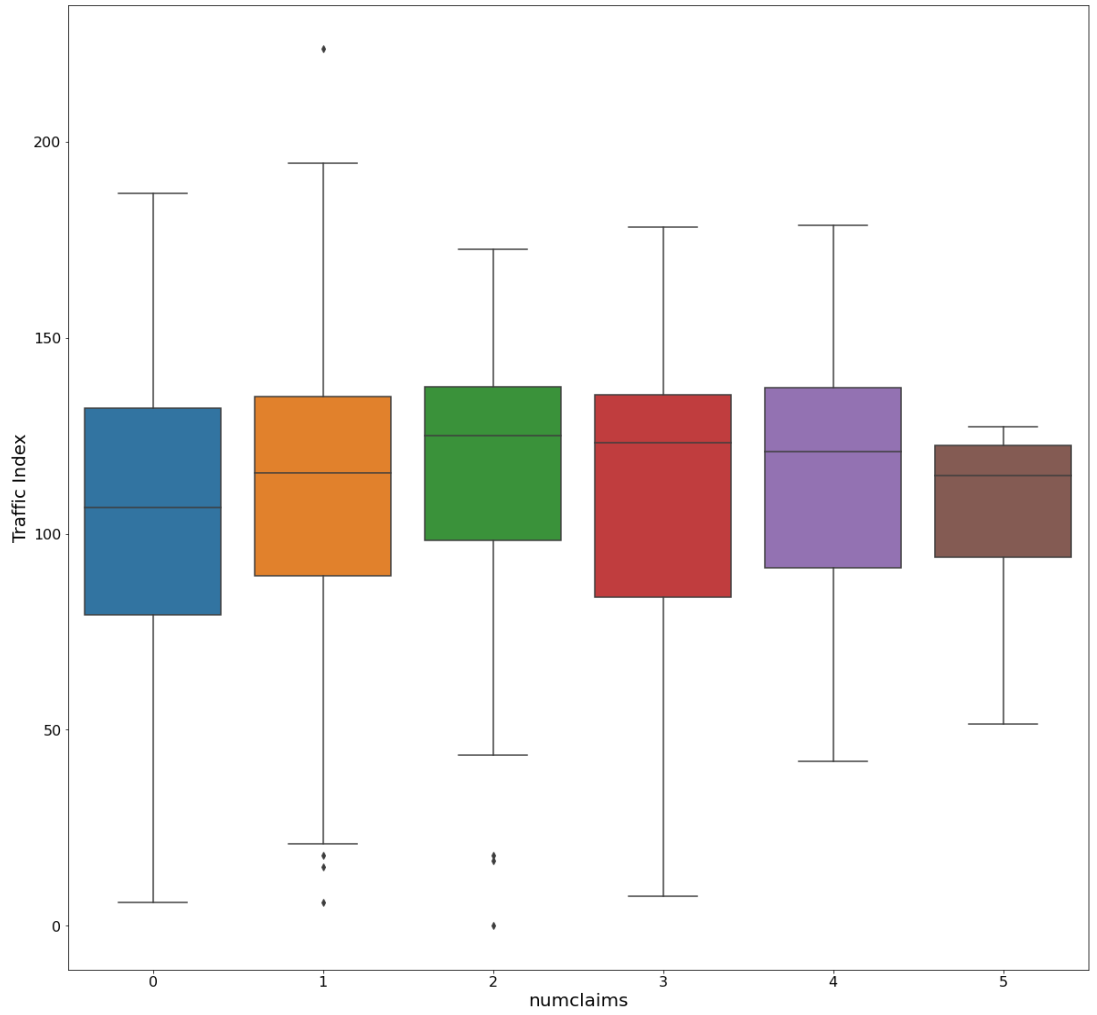
Credit Scores

- The drivers who are predicted to have no claims had the highest median credit score closely followed by the group with one predicted claim.



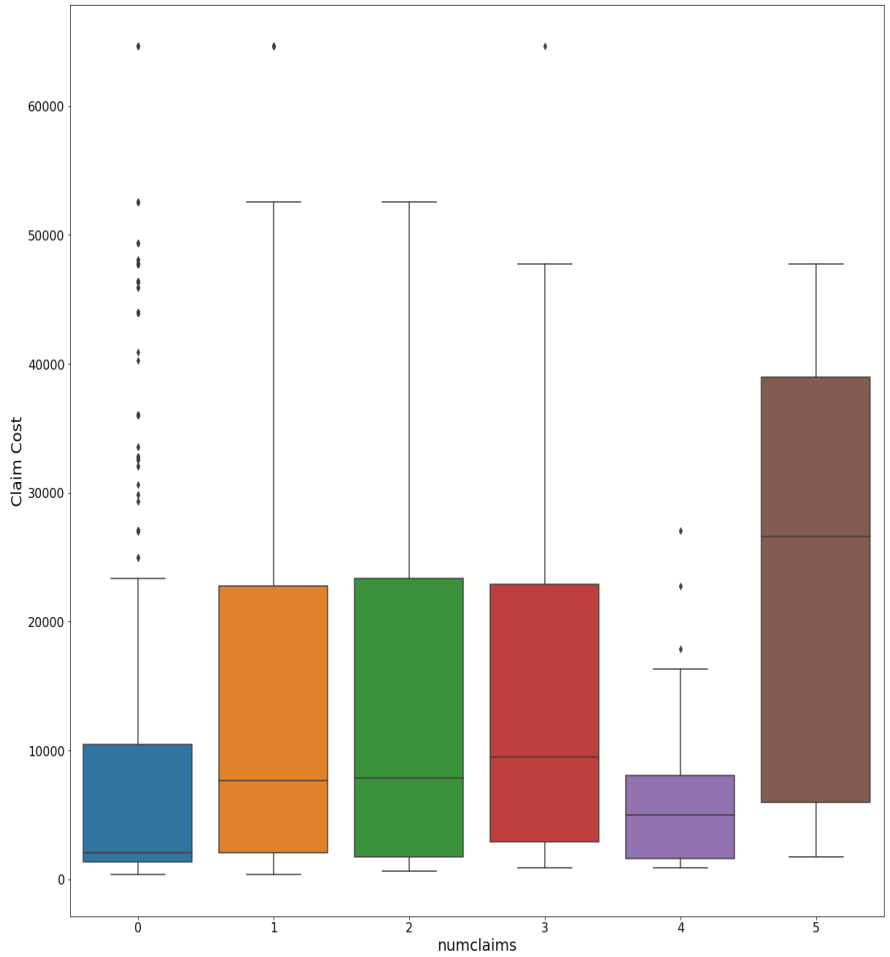
Traffic Index

- Traffic index had almost no impact on number of claims



Claims Cost

- Number of claims highly influenced cost of claims.



Business Value

- Since the median cost of claims of those with 0-1 claims is much lower than the median cost of those with 2-3 claims, and the median cost of those with 2-3 claims is much lower than those with 4-5 claims on our test data, releasing an ad campaign to those with 2-5 claims would be a fiscal disaster to the company.
- This plan shifts the ads to the least risky customers which would greatly increase the profit of this company when you compare that to an ad that randomly targets customers.

Summary and Recommendation

- To recap, credit score and number of predicted claims had the highest impact on the cost of claims. As a result, I think it would be in the company's best interest to only advertise to those who are predicted to have no claims or only one claim.
- Questions?

Missing Values

- I was able to reduce the number of missing rows considerably for both the training data set and the testing data set by determining the bounds of the age category column, then creating a new age category column by binning the date of birth column according to the bounds I found in the original age category column. I was able to do this by separating the date of birth into year, date, and month columns. This allowed me to bin the age category by year. I then deleted the original age category column. Once I did that, I was able to I increase the number of usable rows in the training data from 49919 rows to 54252 rows, and increase the number of usable rows in the testing data from 6124 rows to 6684 rows

Approximating Error

- I approximated error using a 75/25 independent test split on all of the data in the project.
- For the classification portion of the project- number of claims- I used F1 scoring in order to determine the accuracies for the models.
- For the regression portion of the project-cost of the claims- I used mean absolute error to determine the best fitting model.

Parameter Selection

- I chose the parameters that would give me the highest F1 scores and the lowest mean absolute error.
- I decided which parameters for each algorithm was best to use by cycling through various parameters and seeing which ones gave the best results(high F1 and low mae)

Feature Selection

- My strategy behind the features I chose to keep for the dataset was to maximize the functionality of the dataset while minimizing the amount of data being removed from the dataset.
- Although I did delete the columns I hot-encoded in order to be able to scale the data, and I deleted the date of birth for the same reason, I was able to maintain that data in a converted state(hot-encoded, split into day month year)
- The only data column that I deleted with no transmission of data was the policy effective date- which I considered to have no effect on the number of claims and cost of claims.

Class Imbalance

- “From a business perspective, the worst outcome is predicting that a potential customer will have no claims, but in actuality will have any amount of costly claims.”
- Per the instructions to overestimate the number of claims when facing a decision, I chose to use the oversampling method on the decision tree classifier, even though it had a slightly lower F1 score than the original decision tree classifier with no sampling or SMOTE involved.
- Undersampling and SMOTE had a lower F1 score than oversampling and did not address the issue of overestimation brought up in the instructions

Model Selection Motivation- Classification

- I chose to use decision tree classifier with oversampling in the classification section of the project for two reasons:
- Decision tree classifier had a higher F1 score than k-nearest neighbors (95% vs. 85%)
- Oversampling addresses the portion of the project that asks the analyst to overestimate the number of claims(see Class Imbalance).

Model Selection Motivation- Regression

- The first thing I did to find the best regression model was I iterated through several combinations of potential ideal candidates by changing the weights, the algorithm, the number of neighbors, and the p value for k-nearest neighbors regressor and changing the criterion, splitter, and minimum samples leaf for decision tree regressor.
- I found that- from the decision tree regressor- the Friedman mean squared error criterion, combined with the “best” splitter, with a min samples leaf of 5 gave the best mean absolute error of 300 compared to that of 562 with the best combination of attributes for the k-nearest neighbors regressor

Questions?