Car Insurance Prediction Model

Mariia Omelchak, Nico Smith, Johnson Tian, Cameron Zaidi

University of California Davis

STA141C

Goals and Motivation

Federal law (U.S.) requires every driver to have car insurance, which protects pedestrians and other motorists not only from accidents, but also natural disasters. Since around 91% of Americans have a driver license[1] and subsequently car insurance, our group was interested in diving deeper into the world of automobile insurance.

In the insurance industry, one of the most important utilizations of predictive models and user data is the construction of a reliable claim likelihood prediction system, which dictates policy pricing and risk assessment on a case-to-case basis. Meaning, that a poorly trained model can have severe financial repercussions for drivers, pedestrians, and insurance companies themselves. With our knowledge from STA 141A-STA141C we decided to take on the challenge of building a claim status prediction system ourselves, using two powerful models to see which one performs the best.

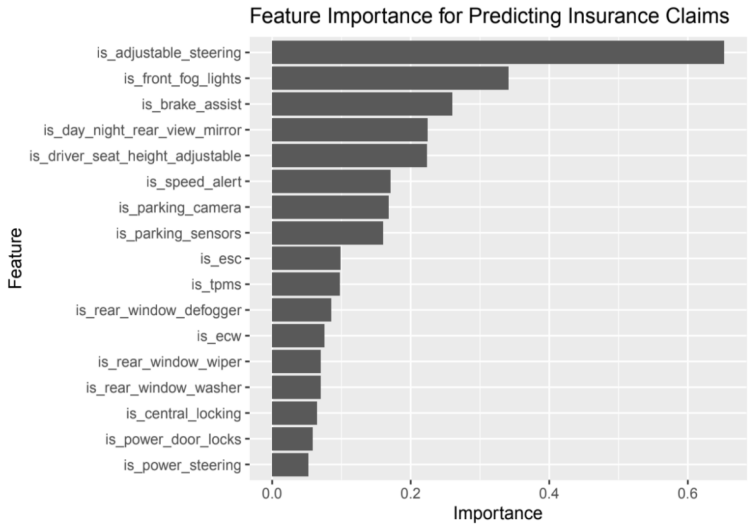
An additional challenge to this problem is that the final claim acceptance call is made by a real person, The Adjuster, who uses risk assessment and pricing prediction models available to them through their specific company policies, as well as human testimonies to decide the drivers' fate. To imitate a human brain, we are using the Neural Network approach, specifically due to its' layering feature which enables hierarchical learning, outputting a more complex and precise model with higher layers. And its competitor in our project is the Random Forests algorithm which imitates many mim-adjusters through the decision trees, and has the capability to highlight or ignore parts of the claims, just like a seasoned adjuster would be making an educated guess. Our focus for this project is auto insurance data exploration and Neural Network vs Random Forest comparison, since our team decided to practice in-depth data analysis rather than optimization.

Description of Data and Exploration

For this project, we utilize historic data on insurance claims, demographic of the policyholders, previous claims, as well as the general vehicle information, to build our claim status prediction model. We found our 689 kB project data on Kaggle, which was posted by Sergey Litvinenko. Our data consists of 27 qualitative variables and 12 quantitative ones, and in order to successfully process and standardize the input for the model, we used data encoding for conversion of categorical variables into numeric ones, as well as signing binary ones with 1s and 0s. To prepare the data, we continued cleaning it by dropping missing values and dropping the 'policy_id' column due to its' lack of relevance. Policy ID is an indexing method used by insurance companies, but it is not necessary for our project, since we are not interested in the individual policyholders, but rather the trends that allow insurance claims to be accepted/denied.

With the data set cleaned, our next step was to identify the most important features for claim status prediction, to give us a closer look inside of the data set. "Importance" of the features is defined through the application of the Random Forest algorithm on the unsupervised clustering (DBSCAN). Further, ranking the features and returning this plot:

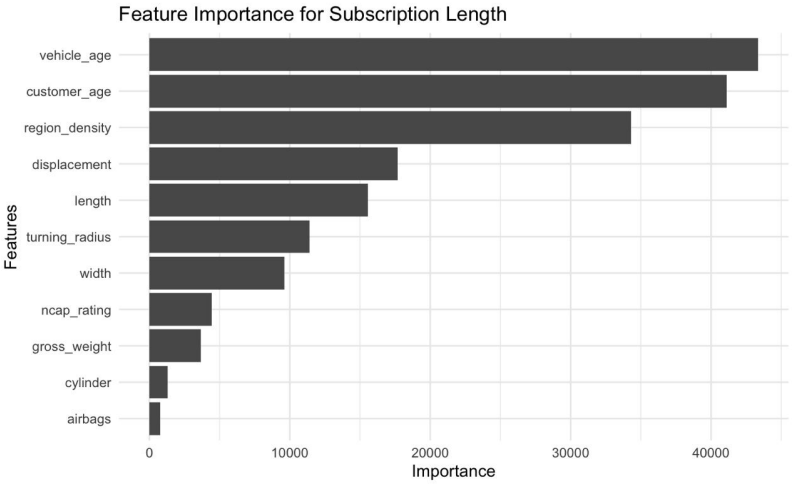
[1] "Number of Licensed Drivers in the US [2024-2025]." Hedges & Company.

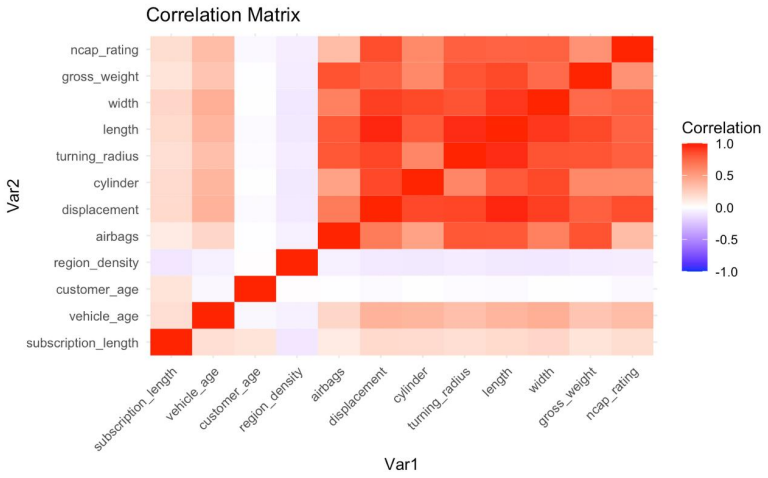


It's evident that by a wide margin, adjustable steering has a crucial impact on the data. This may be due to the fact that newer models and higher end sedans tend to have this feature, policyholders of which have to pay more for their car insurance.

It's important to note that top ranked features such as front fog lights, brake assistance, or day/ night rear view mirrors deal with the safety, visibility, and comfort of the driver and can generally be considered to be "Advanced Safety and Convenience Features".

In our exploration of the data we found that subscription length of the policyholders is another important feature of the claims prediction projess. To dig deeper into this discovery, we created an additional feature importance visualization using Random Forest and a heat map correlation matrix of the variables seeded out by the previous model:





Through the ranking of the subscription length impactful variables, it's clear that customers of different age groups may have varying subscription habits. And older customers may be more conservative and inclined towards longer subscription terms. They are also more likely to own older vehicles, which may be because they change vehicles less frequently or are more inclined to buy used vehicles. Customers living in low-density areas may choose longer subscription terms because there are fewer public transportation options in these regions.Vehicles with larger engine displacements might be associated with longer subscription terms, as these vehicles typically have better performance, require more maintenance, and are suitable for long-term use. They also generally have larger body dimensions, which is consistent with the physical characteristics of the vehicle. Additionally, these vehicles are generally heavier.

Longer subscription lengths may be associated with older customers and older vehicles. This may be because older customers are more likely to choose long-term subscriptions, or older vehicles require more frequent maintenance, resulting in longer subscription periods. And as discussed in the first plot, "Advanced Safety and Convenience Features" also have an effect on the final claim status prediction. In short, older, loyal customers and policy holders with higher premiums (more expensive vehicles) are important demographics to consider in the claim prediction model.

Methods

With clean, ready data, our first goal was to alleviate the class imbalance through SMOTE (Synthetic Minority Oversampling Technique), and split the data into 60% training and 40% testing sets. While researching auto insurance prediction models, we discovered that a prominent split used by many statisticians for the Neural Network approach is the 80/20 split. However, due to our emphasis on the data exploration aspect for this project, we opted for a 60/40 split instead. This was due to the class imbalance of the car insurance industry, as they decline most claims in order to flip a profit.

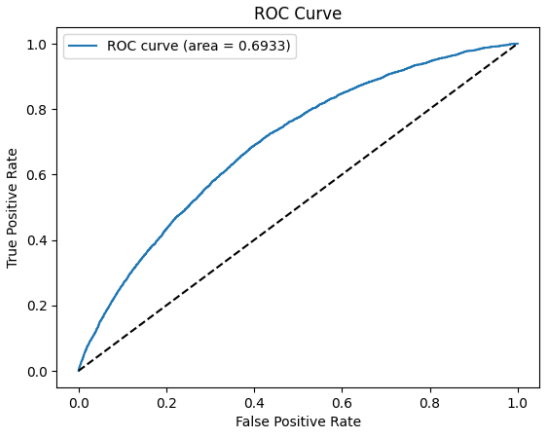
Jumping into the application of the Neural Network application, it's hard to pin down it's inner working as a lot goes on behind the closed doors with this model, some say it is somewhat of a 'black box', so for information of the process we will look at the reported hyperparameters to understand the process better. To begin, we defined a Sequential model architecture with:

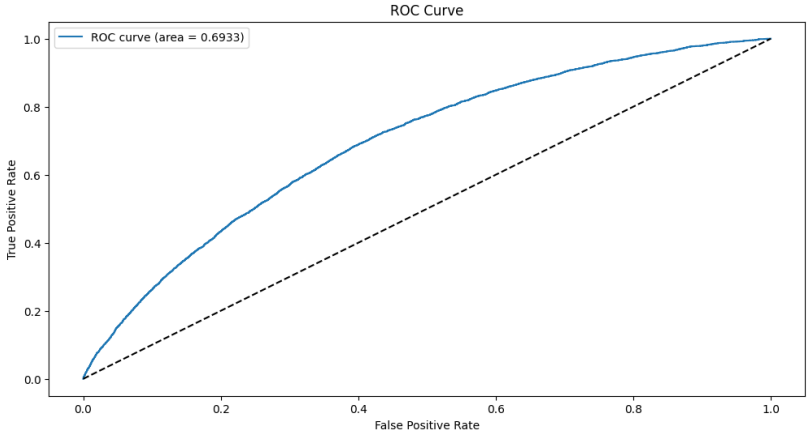
input layer with neurons number of neurons and activation function; batch Normalization and Dropout layers for regularization. We set the Hidden layer to half the number of neurons and used the activation function. As for the Output layer, we used sigmoid activation function for it. Then compiling the model with optimizer, binary_crossentropy loss, and accuracy metrics, we trained it with the best hyperparameters on the training data. These are the hyperparameters we got thought he use of Hyperopt for hyperparameter optimization:

- Activation: relu
- Batch size: 113
- Dropout rate: 0.24
- Epochs: 36
- Neurons: 191
- Optimizer: adam

[2] V. R. Joseph, Optimal ratio for data splitting, Stat. Anal. Data Min.: ASA Data Sci. J.. 15 (2022), 531–538.

Both relu and adam are baisc models for Neural Network training, so the process does not seem to have an extremely tailored or complex approach with our data. We repeated the process 3 times, due to the hierarchical learning ability that Neural Network has, sharpening the predictive power of the model with every layer. This was the returned ROC curve for the first layer:





The accuracy of this beginning layer (on the left) is 69.33% and this is the same accuracy for the second layer (right), which is not ideal, considering that a Neural Network on average has a minimum accuracy level of 70%. We continued adding layers to our model in order to improve the accuracy of the prediction.

Next, we applied the Random Forest algorithm, leveraging k-fold cross-validation to enhance model robustness. To define the architecture of the tree, we included 100 decision trees in the forest (n_estimators), used the square root of the number of features for max_features,

optimized depth (max_depth), minimum samples for split (min_samples_split), and minimum samples for leaf (min_samples_leaf). Next, feature pruning was performed using prior feature importance analysis and we employed GridSearchCV to identify the best hyperparameters for fine-tuning our model. This was the output for the best hyperparameters, similar to that we got for the Neural Network model:

Best parameters	Optimal values
n_estimators: 100	max_depth
max_features: sqrt	min_samples_split
	min_samples_leaf

This comprehensive approach contrasts with simpler methods seen in other Kaggle projects where single train-test splits and less rigorous handling of class imbalances may lead to less reliable results. Our methodical data preprocessing of strategic class imbalance handling and thorough hyperparameter tuning distinguished our approach from typical methods used in other insurance claim prediction projects.

Results

The results of our car insurance claim prediction project demonstrated the efficacy of our comprehensive approach. The Random Forest model achieved a robust performance reflected in a high ROC-AUC score of 0.6979 Additionally, we achieved an F1-score of 0.89, indicating a balanced performance in terms of precision and recall. These results underscore the value of our extensive data preprocessing, strategic handling of class imbalances through SMOTE, and meticulous hyperparameter tuning via GridSearchCV and Hyperopt. The Random Forest model particularly excelled, benefiting from feature pruning and k-fold cross-validation, which ensured the model's robustness and generalizability. Our approach outperformed many simpler methodologies typically seen in other Kaggle projects and insurance claim studies, highlighting the effectiveness of our rigorous process and thorough analysis.

Discussion

Our approach to predicting car insurance claims showcased several strengths including robust data preprocessing, effective class imbalance handling through SMOTE, and meticulous hyperparameter tuning with GridSearchCV and Hyperopt. These steps ensured high model performance, as evidenced by our Random Forest model's ROC-AUC score of 0.6979 and the Neural Network model's overall accuracy and F1-score of 0.89. The use of k-fold cross-validation further strengthened our results by minimizing overfitting and enhancing generalizability. Our approach also had some limitations. The choice of a 60/40 data split, while beneficial for extensive data exploration, might have impacted the model's ability to generalize

as effectively as the commonly used 80/20 split. Additionally, the models could be improved by exploring more advanced feature engineering techniques, integrating additional ensemble methods, and incorporating deeper neural network architectures. Future work could focus on these areas, as well as experimenting with other oversampling and undersampling techniques to handle class imbalances more effectively. Enhancements in these aspects could further boost the predictive accuracy and robustness of our models.

Conclusion

Our car insurance claim prediction project demonstrated the effectiveness of combining robust data preprocessing, strategic handling of class imbalances, and meticulous hyperparameter tuning. The Random Forest model and Neural Network model both achieved high performance metrics, with the latter attaining an overall accuracy and F1-score of 0.89. Some important features that we discovered are "Advanced Safety and Convenience Features" and customer subscription length)loyalty to the insurance company), which are likely to have a positive effect on the Claim Adjuster's decision. Our use of k-fold cross-validation ensured the reliability and generalizability of our models. Despite the strengths of our approach, there are areas for improvement, such as exploring more advanced feature engineering techniques and additional ensemble methods. Future work focusing on these enhancements could further increase the predictive accuracy and robustness of our models, solidifying their utility in real-world insurance claim prediction scenarios.

CODE:

All code will be attach in the submission
Graphs: STA_141C_FP.rmd
More Graphs: STA_141C_FP2
Actual Modeling Code: 141C_final.ipynb
Data: https://www.kaggle.com/datasets/litvinenko630/insurance-claims

REFERENCES:

"Number of Licensed Drivers in the US [2024-2025]." Hedges & Company., 7 Mar. 2024,hedgescompany.com/blog/2024/01/number-of-licensed-drivers-us/#:~:text=We're%20now%20projecting%20over,impact%20on%20total%20licensed%20drivers.

V. R. Joseph, Optimal ratio for data splitting, Stat. Anal. Data Min.: ASA Data Sci. J.. 15 (2022), 531–538. https://doi.org/10.1002/sam.11583