Insurance Claims Prediction

Problem

Model the expected claim amount per policyholder and per year based on customers' risk characteristics.

Solution

After gaining some insights through EDA, the forecast problem is tackled in two steps. In the first one, an XGBoost logistic classification model will predict if the claim amount is going to be zero (0) or higher (1). In the second step, an XGBoost regression model will improve the strictly positive predictions.

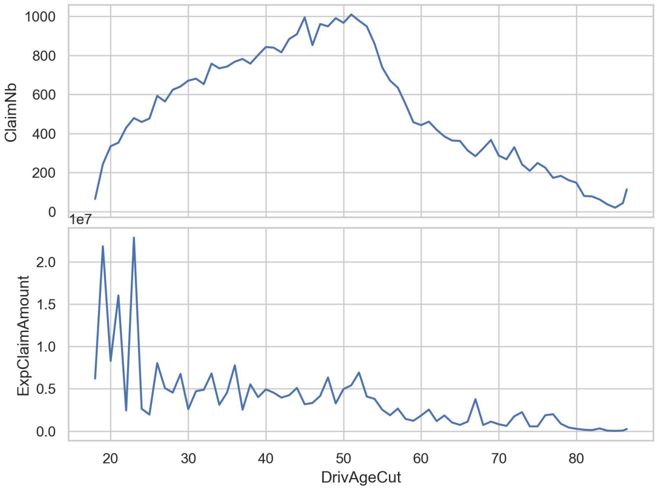
Key Insights

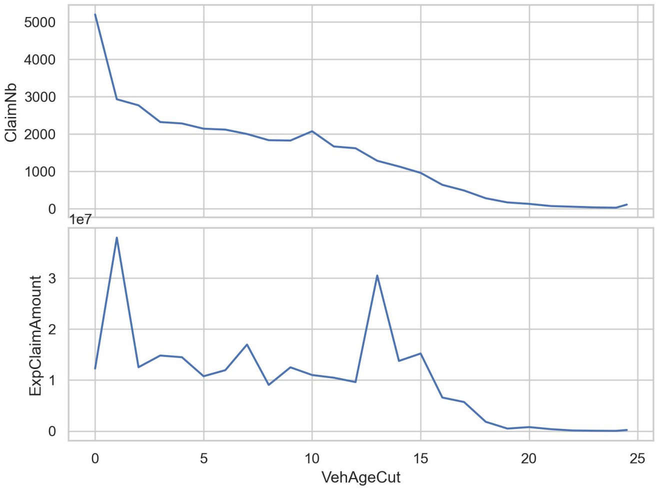
1. In some areas and regions, the number of claims and the total claimed amount is higher than in others.
2. The smaller the bonus-malus penalty, the higher the number of claims. This does not apply to the total claimed amount.
3. The higher the population density, the higher the number of claims. This does not apply to the total claimed amount.
4. The number of claims increases with driver age, up to the age of 50, after which it decreases.
5. The total claimed amount is highest among 18- to 25-year-old. It remains somewhat constant for older people.
6. The number of claims decreases the older the vehicles are. The total claimed amount also decreases with vehicle age, with peaks at 1 year old vehicles and between 12-15 years old vehicles.
7. The best logistic XGBoost model does well at classifying zero values, but has problems with positive values. It has the following metric: 93% Precision, 78% Accuracy, 78% Recall, 0.53 ROC-AUC score.
8. The best XGBoost regression model performs worse than a mean model. A different approach is needed.

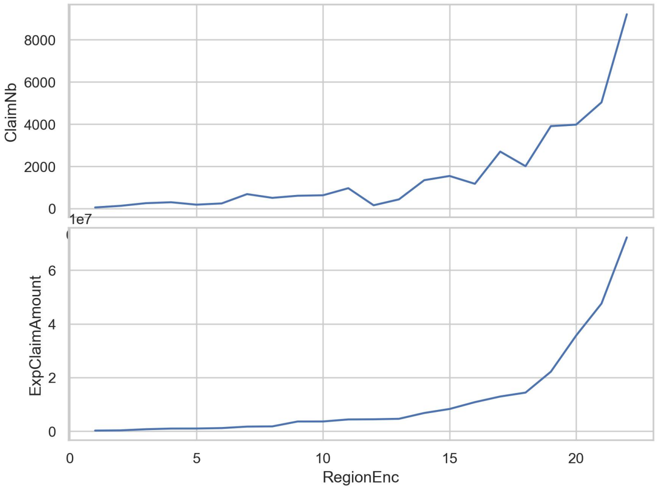
07.03.2024, V. Ganci, GitHub Portfolio

Next Steps

1. Apply oversampling techniques to reduce the class imbalance (null, non-null X).
2. Apply hyperparameter tuning to the models via GridSearchCV to improve the models' metrics.
3. Test the performance of a Neural Network model.


ClaimNb
1e7
ExpClaimAmount
DrivAgeCut


ClaimNb
1e7
ExpClaimAmount
VehAgeCut


ClaimNb
1e7
ExpClaimAmount
RegionEnc