Porto Seguro's Safe Driver Prediction

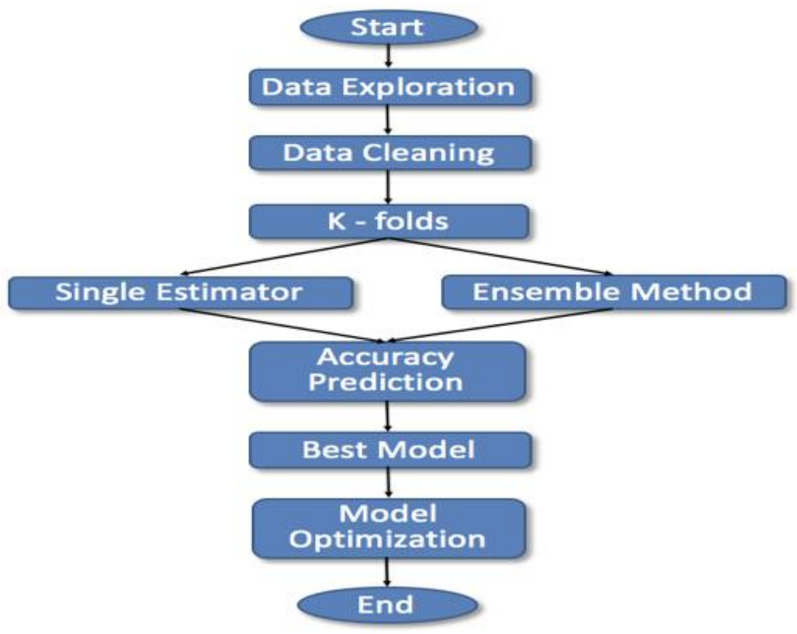
Bing Mei, Ke Cao, Zhenyu Kang
Instructor: Professor Chris Asakiewicz

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction: Problem

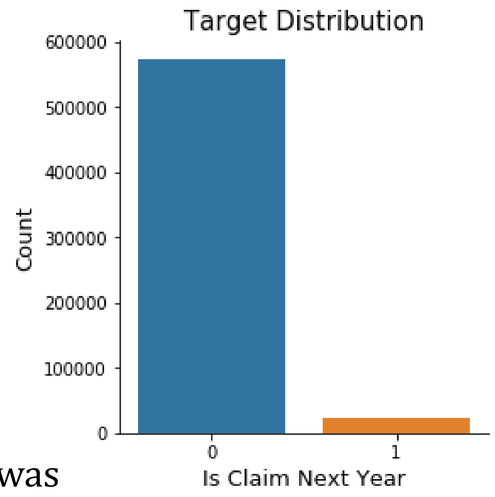
- Porto Seguro is the third largest insurance company in Brazil. In this competition, we will build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year.
- A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.
- The result is evaluated using the Normalized Gini Coefficient.

Flowchart

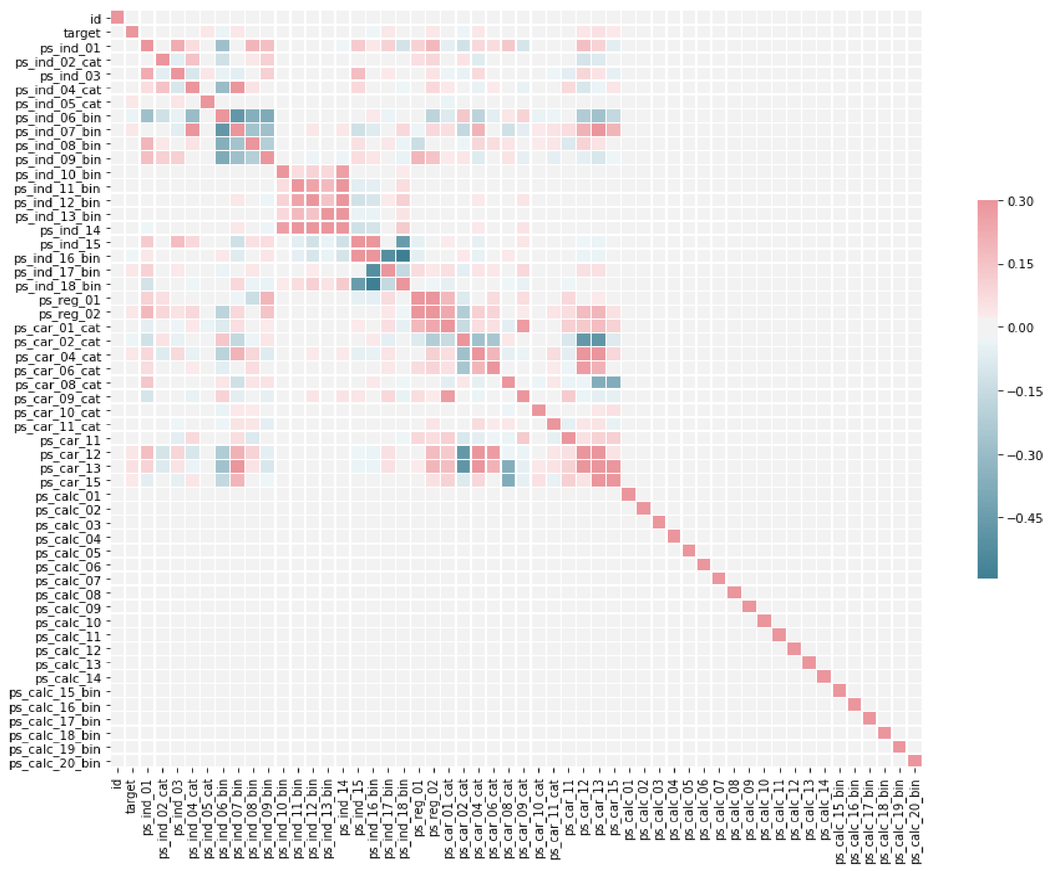


Data Exploration

- Training data set has 595212 entries and 58 variables, the target columns signifies whether or not a claim was filed for that policy holder.
- We have a highly imbalanced target distribution.
- Values of -1 indicate that the feature was missing from the observation.
- The correlation of all variables are shown below.



Correlation Matrix



- ps_calc_* features are not related to target at all.
- Removing them would prevent the curse of dimensionality.

Data Preparation & Modeling

Data Preparation

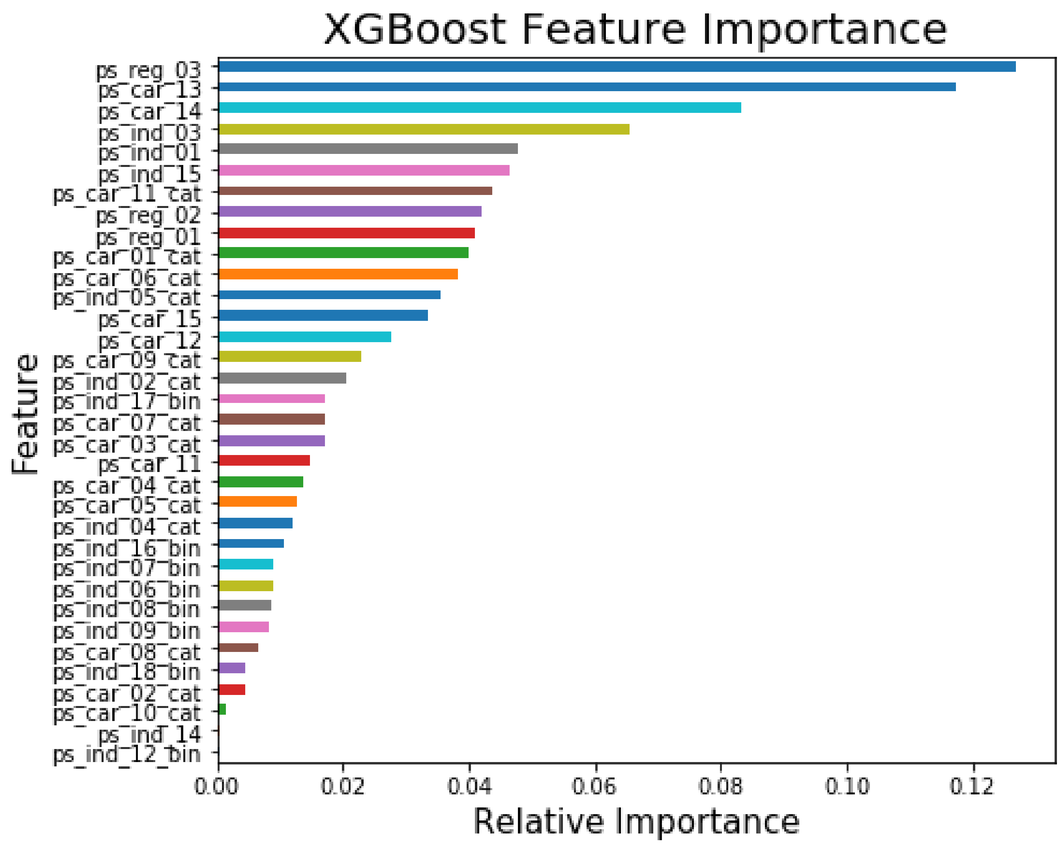
- For missing values, leave -1 in the categorical features and use mean values to replace them in the numerical features.
- Remove the ps_calc_* features to prevent the curse of dimensionality.
- One Hot Encoding for categorical features.

Modeling

- Start modeling with K-fold cross-validation method. (k = 5 here).
- Define the Normalized Gini Coefficient function.
- Select single model estimator (Logistic Regression) and ensemble method (XGBoost).
- Submit the results to Kaggle.com to compare their performances.

Results

- The best score of Logistic Regression is 0.27340.
- The XGBoost model performed significantly better by the final score of 0.28864.
- Our final ranking is 1565th of 5170 (Top 31%).
- The Chart shows the importance of each feature.



Conclusion

- The Ensemble Model performances better than Single Model.
- For the anonymous data, do not drop any data easily.
- The method of how to process missing values will affect the result but not much.
- Some of features which have relative greater importance are continuing features. For example, ps_reg_01, ps_reg_02, ps_reg_03, ps_car_13, ps_car_14. We can construct some combination features for future works.
- We can also eliminate some features with very low performance.