Group 5:
The process for designing our optimal model

Andrew Fraser, Catherine Zheng, Kira Kim, Pearl Linyue Li and Stepan Mincev,

kaggle Competition

Agenda Layout

- Problem at Hand
- Description of Data
- Data Analysis
- Algorithms and Learning Method
- Model Performance
- Concluding Remarks

Problem at hand

Objective - Develop a model that would aid in identifying the risk of a claim by an individual.

How to achieve this:

Use a series of specific variables, that would be available as part of an insurance application, to train a model that would predict probability of an individual making a claim.

Problem at hand

How does predicting this aid the company?

The application of the model would enable the company to set competitive insurance rates for those who are not likely to make claims while providing appropriate rates for those who are.

This would lead to:

- Smaller losses due to paying out claims.
- Increased income from increased market share

Description of data

In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc). In addition, feature names include the postfix bin to indicate binary features and cat to indicate categorical features. Features without these designations are either continuous or ordinal. Values of -1 indicate that the feature was missing from the observation. The target columns signifies whether or not a claim was filed for that policy holder.

https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data

Data Analysis

Feature Engineering

- Pairwise relationship of features
- Relative contribution of values to the target prediction to prioritise covariates with the highest discriminative ability (filtering out *_calc* variables)

Data Pre-processing

- Conversion and factorisation of categorical values
- Exploration of imputation methods for missing values including (mode/mean, omission, predictive imputation, categorisation)

Algorithms and Methodologies:

Algorithm that were tested:

- XGBoost
- GLM
- Neural Network
- Random Forest

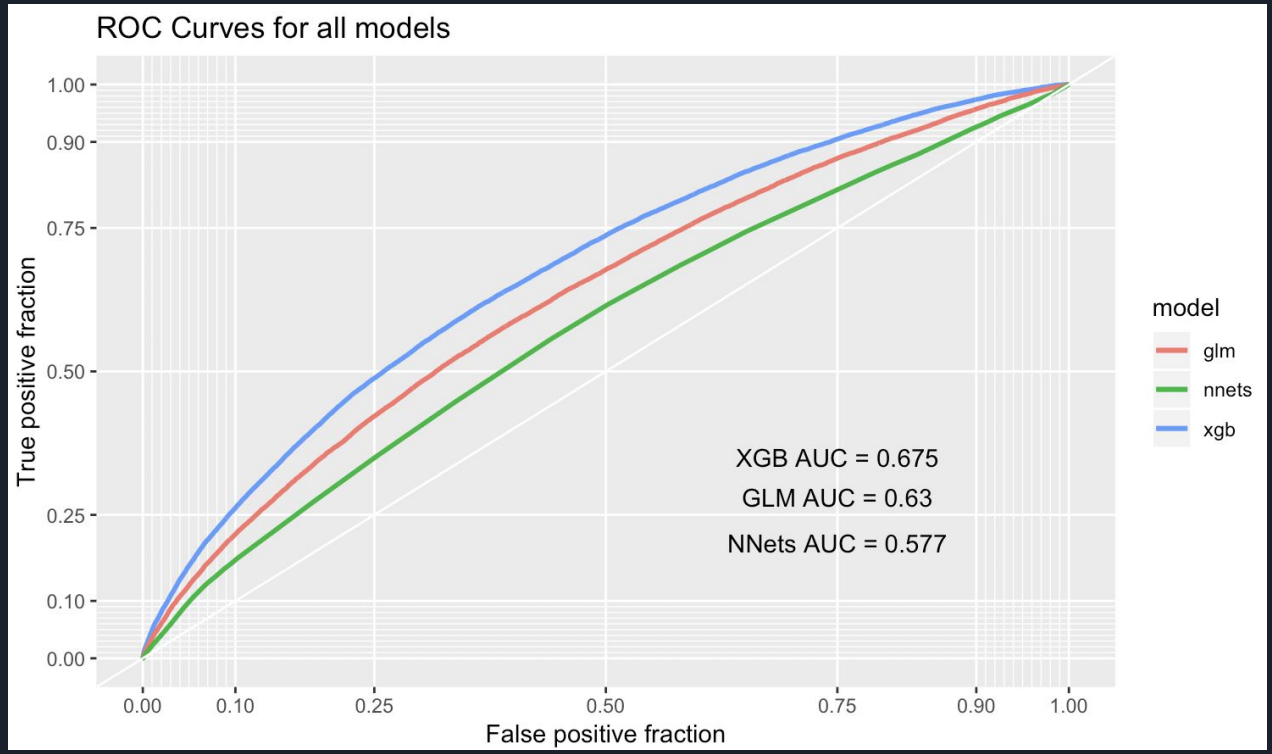
Training Methodology:

- Basic model with default configuration
- Extended model with incremental model complexity
- Use of GridSearch for hyper-parameter tuning to attain the optimal model
- Cross Validation at training

Model Performance

Model	Public Score	Private Score
Xgboost	0.28243	0.28641
GLM	0.23079	0.23761
Neural Nets	0.13324	0.14371

Model Performance



Concluding Remarks

Chosen model:

- XGBoost

Most significant variables for predicting whether a person will make a claim:

- ps_car_13,
- ps_reg_03
- ps_ind_05

References

Steering Wheel of Fortune - Porto Seguro EDA (2018) Available at: https://www.kaggle.com/headsortails/steering-wheel-of-fortune-porto-seguro-eda/