Project Proposal
Data Science Practicum
Spring 2019

Customer Transaction Prediction for Santander Bank (Kaggle competition)

March 8th, 2019

Team name: *Squadron*

Team Members:

- ★ Denish Khetan; denish.khetan@uga.edu, LinkedIn
- ★ Mohammadreza Iman; mi44512@uga.com, LinkedIn

Project description:

Kaggle competitions: "Kaggle Competitions are designed to provide challenges for competitors at all different stages of their machine learning careers. As a result, they are very diverse, with a range of broad types. Featured competitions are the types of competitions that Kaggle is probably best known for. These are full-scale machine learning challenges which pose difficult, generally commercially-purposed prediction problems."

Kaggle Customer Transaction Prediction for Santander Bank challenge: "In this challenge, we invite Kagglers to help us identify which customers will make a specific transaction in the future (classification), irrespective of the amount of money transacted. The data provided for this competition has the same structure as the real data we have available to solve this problem."

- ➢ Dataset: Anonymized dataset containing numeric feature variables, the binary target column (classification), and a string ID_code column.
- ➢ Training dataset: 200,000 samples consist of 201 features and the target
- ➢ Test dataset: 200,000 instances consist of same 201 features of the training set

"Submissions are **evaluated** on area under the ROC curve between the predicted probability and the observed target."

Source: https://www.kaggle.com/c/santander-customer-transaction-prediction

Preliminary plan and steps:

- Arranging the GitHub repository and documentation platform
- Data analyzation and understanding the features, documenting the features' statistical profiles
- Data preprocessing (feature selection, normalization, etc.) if necessary
- Applying mathematical models, e.g. regression-based methods, k-nearest, and SVM
- Applying Trees methods (J48, Random Forest, etc.)
- Applying the neural network based methods, e.g. ANN and DNN
- Applying the sum-product networks, extra (if we have enough time)
- Documenting the process
- Documenting details of the best found model
- Preparing the presentation

Schedule:

- Setting up the development environment by March 20
- Data analyzation and preprocessing by March 23
- Applying the models and find the best model and fine-tune it by April 5
- Final submission to the competition's leaderboard by April 10 (End Date of competition: April 10, 2019 11:59 PM UTC)
- Finalizing the documentation by April 17
- Finalizing the presentation by April 28