**Classification Project Write-up**

**Predicting default rate of Home Credit Loan Customer**

**Abstract**

The goal of this project was to use classification models to predict whether a client can repay the loan for Home Credit bank in order to help improve the approving process of client application. I worked with data provided by Kaggle which contains demographic features and credit features of each applicants and build a classification model on the data. After refining a model, I put together a PowerPoint to summarize my findings and present to my audience.

**Design**

This project originates from the Kaggle challenge: Can you predict how capable each applicant is of repaying a loan? The data is provided by [Kaggle (Links to an external site.)](https://www.kaggle.com/competitions/home-credit-default-risk/overview) and the target variable is the status of the clients’ default. Classifying statuses accurately via machine learning models would enable the home credit bank to take action to improve operations and target suspicious clients more efficiently in the reviewing process.

**Data**

The dataset contains 300k+ client’s info with 122 features for each, 52 of which are categorical. A few feature highlights include client’s credit score, age, annuity rate, family size, educational experience. Nearly a fourth of the categorical features could be grouped into more general categories, and an in-depth analysis of 20 of them was undertaken to inform baseline models and feature engineering.

**Algorithms**

*Feature Engineering*

1. Changing categorical feature ‘educational level’ to ordinal int type
2. Converting categorical features to binary dummy variables
3. Combining particular dummies and ranges of numeric features to highlight strong signals during EDA
4. Changing daily variables into yearly variables by dividing by 365
5. Creating a 500 nearest neighbor feature to compute client’s distance (two most important feature’s distance) and formulate an important feature

*Models*

Logistic regression, random forest, and gradient boost classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

*Model Evaluation and Selection*

The entire training dataset of 307,507 records was split into 80/20 train vs. test, and all scores reported below were calculated with 3-fold cross validation on the training portion only. Predictions on the 20% test set were limited to the very end, so this split was only used and scores seen just once.

The official metric for loan default rate was classification rate (accuracy); however, class weights were included to improve performance against imbalance dataset (10% minority is default entries).

**Final random forest 3-fold CV scores:** 20 features with class weights

* 3 candidate models ROC-AUC score:
* Logistic regression: 0.58
* Random forest (selected): 0.72
* Gradient boost: 0.71

**Tools**

* Numpy and Pandas for data manipulation
* Scikit-learn for modeling
* Matplotlib and Seaborn for plotting

**Communication**

A slide is put together to summarize the findings and communicate with audience.