**Credit No-hit Project**

**BU - Capstone Project Summary**

*Team: Yulong Gong, Muyan Xie, Yichi Zhang, Yangyang Zhou*

**Problem Statement & Goal**

Nowadays, credit status becomes an essential part of both personal life and business settings. For business entities, whether a consumer has a credit record or not can be a good indicator to predict potential gain or loss. For this project, we worked with a home insurance company using data provided to help them to filter potential customers, reduce potential losses, and eliminate ~ 3 months of the credit checking process.

**Methodology Overview**

* Data Preparation:
  + Data cleaning & data integrity checking (missing values & outliers)
  + Data Visualization
* Model development
  + Plan brainstorming (dataset with/without time info)
  + Sample design (target design, stratification, and resampling)
  + Feature reduction (correlation analysis, feature importance analysis)
  + Model selection (RandomForest, ExtraTrees, XGBoosting, AdaBoost, DecisionTree, KNeighbors, LogisticRegression, Bernoulli, RidgeRegression, Gaussian, Perceptron)
  + Model combination (stacking)
  + Model evaluation & tuning (cross-check & parameter tuning)

**Project Outcome**

The model will give the probability of credit no-hit for a potential customer with limited information. The top 10% riskiest portion identified by the model captured 54.77% of all ‘no-hit’. It could help the company decide whether they want to develop a business relationship with this customer. It will take about one year “baking period” for the company to evaluate how well this model performed.

**Tools & Technology Used**

Programming: Python (desktop & Google Cloud Platform, Google Colab), Jupyter Notebook, GitHub for version control.

Data management: Google Drive, Google Cloud Platform.

**Key Learnings**

* Communication:
  + Communicated with supervisors in a professional setting.
  + Coordinated with team members and developed strong team morale.
* Programming：
  + Learned how to clean data and do feature engineering on a large size real-world dataset to help achieve certain business goals with SQL and Python.
  + Handled imbalanced data to improve model performance for binary classification problems.
  + Got used to developing self-defined functions to develop a data preparation pipeline which increased the efficiency of model training and usability for the long term.