**Capstone Project**

**Project proposal**

**GROUP NAME**

|  |
| --- |
| DropNA |

**STUDENT INFO**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Background** | **Target Industry** |
| James Kim | Applied Math | Finance/FinTech |
| Huimin Ou | Engineering | No target industry, open to diverse opportunitie |
| HanXiao (Mia) Zhang | Finance | Finance |
| Liyi (Lily) | Engineering | Finance/FinTech/Tech |

**GROUP STRUCTURE**

|  |  |  |
| --- | --- | --- |
| **Student Name** | **Data Science** | **Project Team** |
| James Kim | - data engineering  - machine learning | - project proposal submission  - presentation |
| Huimin Ou | - data engineering  - exploratory data analysis | - time management  - technical lead |
| HanXiao (Mia) Zhang | - EDA and insights  - machine learning | - execute project plan  - presentation |
| Liyi (Lily) Kuo | - data visualization  - data engineering | - project proposal submission  - execute project plan |

**PROJECT OVERVIEW**

|  |
| --- |
| **Overview:**  Lending club’s stock performance has crashed 50% since the 2016 scandal and has unable to recover since the crash. Based on the perspective of a team of data science consultants, we seek to provide an alternative rating system for the borrowers and provide insights on the probable risk of opportunity lost and loan defaults.  **Objective:**   * Identify important features that are significant to loan grade/sub grade * Provide an alternative grading system to each individual loan * Provide Survival Time Analysis to client   **Measure of Success:**   * Identify the features that are significant in determining the loan grades/sub grades * Predict the loan performance in term of loan maturity and PnL * Identify the risk of losing the opportunity cost from the investor’s perspective * Identify the loans that are highly likely to default post issuance * Build models to predict loan duration for charged off/default loans and for good loans which terminate with 'fully paid' status. * Build models to predict total profits/losses (or total principal/interests * Show evidences to link the loan-duration and the PnL, which may justify training a joint model predicting multi-regression-targets. |

**SCOPE OF APPLICATION**

|  |
| --- |
| **Population**:  Process and model will be applied to new customers requesting for a loan, and customers who were granted a loan after 2015  **Timeframe**:  The Lending Club has 10 years of data for this business, but the portfolio structure and marketing has critically changed throughout the years, especially after 2016. Our model will be constructed to adjust to the changes  **Target variable:**  Customers paying off their full loans prior to maturity  Customers default on their loans prior and post due date  Questions to answer with business and risk:  - What is the lost in company profit based due to opportunity lost?  - What is the lost in company profit due to customer defaults? |

**DATASETS**

|  |
| --- |
| **Dataset 1: Loan Accepted Data**  - Internal system, approved loans  - Past 12 years of data [from 2007-2018]  - Variables: request id, loan amount, term, interest rate, FICO score, …  **Dataset 2: Loan Rejected Data**  - Internal system, rejected loans  - Past 12 years of data [from 2007-2018]  - variables: amount requested, application date, loan title, risk score…  **Dataset 3: Bureau data**  - external data  - National bank benchmark interest rate |

**DATA TREATMENT AND ANALYSIS**

|  |
| --- |
| **Data preparation**  - merge 2007-2016 and 2017-2018 datasets  - EDA on both accepted and rejected loan  - Feature engineering on the merged datasets [find missingness, inputing,…]  **Target variable**  - loan interest rates  - loan grades and subgrades  **Tools**  - Data extraction in Python  - Data preparation in Python  - Feature classification via PCA, Naïve Bayes, and GBM  - Logistic regression model development in Python  - Random forest model development in Python  - GBM model development in Python  - Combination of models are stack and performance of the stacked models are then examining against non-stacked models  **Analysis**  - Classification model  - Prediction model: test logistic regression as baseline and random forest |

**PROJECT TIMELINE**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **November** | | | | | | **December** | | | | | | | | | | | |
| 25 | 26 | 27 | 28 | 29 | 30 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Kick off** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Project declaration* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data extraction (ETL) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Explore data analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train Naïve Bayes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test Naïve Bayes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalize results |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Presentation prep |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Buffer |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 2** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Delivery** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |