**Capstone Project**

**Project proposal**

**1.** **Group description**

**1.1.** Group name

|  |
| --- |
| dotFit |

**1.2.** Students names, background and target industry if any

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Name | Background | Target Industry | | Jessica Joy | BS Financial Economics | Undecided | | David Kim | Business development | Tech or Finance | | Eugene Ng | Financial services  Risk management | Finance related (but flexible) | | Darish Sakeesing | Information systems | Tech or Finance | | David Wasserman | Credit Risk and Technology | Fintech | |

**1.3.** Group structure: roles and responsibilities

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Name | Data Science | Project Team | | Jessica Joy | EDA, visualization, ML | Presentation | | David Kim | EDA, visualization, ML | Classification ML  Presentation | | Eugene Ng | EDA, visualization, ML | Project proposal submission  Time management  Presentation | | Darish Sakeesing | EDA, visualization, ML | Technical lead  Presentation | | David Wasserman | EDA, visualization, ML | Technical lead  Presentation | |

**2. Why** do we want to develop a data science project?

**2.1 Objective**: what problem do you want to solve? What questions are you trying to answer? How will you **measure the success** of your analysis from a business/user perspective?

|  |
| --- |
| **Context**: Investors who participate in the Lending Club platform are presented with a multitude of investment options – an extremely diverse set borrowers each with unique loan needs, credit profiles and income levels. The goal of any investor is to maximize returns while minimizing losses, so our data science team was contacted by a group of investors to assist in the investment selection process.  **Objective:**   * Provide valuable EDA and visualization on the Lending Club portfolio of loans answering key questions such as:   + What is the default rate? What are the defaults per loan grade?   + What are the most common loan purposes?   + What are the interest rates for each loan grade (time series and snapshot)   + What types (loan purpose, profession, term, state, etc.) of loans tend to prepay or default?   + Which states, professions, loan durations, income levels, etc. are most common?   + Segmentation by FICO scores of the borrowers   + **The above is a sample of the many EDA questions that will be explored and answered so that the investor group will have a deep understanding of the Lending Club product before committing capital.** * Build classification machine learning (ML) model(s) that the investor group can use to predict two key risk events: loan default or loan prepayment * Build regression ML model(s) to predict   + Duration of loans   + Profit/loss of loans * Neural network to predict loan performance   **Measuring Success:**   * Provided with valuable insights and ML tools, investors will be able to confidently navigate the Lending Club portfolio selecting the best loans that maximize returns given their specific risk appetite. |

**2.2.** **Scope** of application: what population and timeframe will your analysis/model be applied to or used for?

|  |
| --- |
| **Population:** Borrows seeking funding in the range of $1,000 - $40,000  **Timeframe**: Lending Club has provided 11 years of data (2007-2018). The platform removed grades F1 to G5 in 2017 due to increased levels of prepayment and delinquency. We may omit all loans of these grades from the analysis as they no longer represent the current the current investment environment.  For similar reasons, we may consider using less data for model training based on findings from the EDA process. For example, the loan portfolio from 2008 could be significantly different from 2018, so training a model that incorporates 2008 data may reduce the model’s predictive / explanatory power.  **Target Variable:**   * Loan default * Loan Prepayment * Loan duration * Loan profit/loss |

**3. How** do you translate the objective and scope in terms of data?

**3.1.** What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

|  |
| --- |
| **Accepted Loan Dataset**:   * All approved and funded loans * 11 years of data. ~2.2 million approved loans * Loan origination variables: loan amount, FICO score, income, term, state, etc. * Loan performance variables: loan status, hardship/workout information, delinquency, etc. |

**3.2.** What **data treatment and analysis** do you plan? Data Aggregation, target variable definition, tools, analysis/machine learning, ...

|  |
| --- |
| **Data Preparation**   * Missing data – analysis and imputation (where applicable) * Removal of loan grades F1 to G5 * Clean-up / clustering of emp\_title field * One hot encoding and label encoding for various categorical variables * Various feature selection and engineering * Pooling and/or time-series   **Tools**   * Dask to manage the large dataset * Data preparation in Python * Regression models in Python or R * Neural Network in Python * Tree models in Python * Boosting/ensemble models in Python * Scikit-Learn’s ‘scikit-survival’ package   **Analysis**   * EDA: univariate and bivariate * Prediction models   + Handling data imbalance: stratification or up-sampling |

**4. Project plan**

Please submit a project plan proposal broken down by a few significant steps. Plan at least three meetings with your stakeholders.

- **Kick-off meeting**: schedule a 30 minute meeting before project declaration in order to approve the project proposal.

- **Milestone 1**: schedule 30 minutes with your stakeholders to present initial results and insights and to validate any assumptions or definitions needed to move forward.

- **Milestone 2**: schedule 30 minutes to go over the final results and proposed presentation before the final presentation in front of the whole team.

If you think additional discussions will be required, feel free to add secondary milestone(s) in your project plan.

**Project plan and schedule examples**

(create and use your own template)

**Kick off**

Project proposal and timeline

**Milestone 1**

Check on progress

**Milestone 2**

Final results Preparation for presentation

**Delivery**

Final presentation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **November** | | | | | | **December** | | | | | | | | | | | |
|  | 25 | 26 | 27 | 28 | 29 | 30 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Kick off** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Project declaration |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ... |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step n |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step n+1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ... |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 2** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Delivery** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |