**Capstone Project**

KCP Investments

**Project proposal**

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

**1.1.** KCP Capital Investments

|  |
| --- |
| Team KCP Investments |

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

|  |
| --- |
| 1. **Jingrou Wei** 2. Background: graduate (M.S. in Mathematics) 3. Target industry: eCommerce and FinTech 4. **Pengyun Liu** 5. Background: business administration/finance 6. Target industry: Finance and FinTech 7. **Peter Liu** 8. Background: credit risk / banking 9. Target industry: finance |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Student** | **Data science** | **Project team** | | Jingrou Wei | - EDA / data engineering  - machine learning  - data visualization | - tech lead  - submission  - presentation  - execute project plan | | Pengyun Liu | - EDA / data engineering  - data visualization  - machine learning | - time management  - presentation  - tech co-lead | | Peter Liu | - EDA / data engineering  - data visualization  - machine learning | - execute project plan  - 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**: we are looking to help individual investors seeking exposure to the higher-yielding consumer loan space to diversify his/her investment portfolio.  **Objective**:  - Predict and assess the level of risk of loan applicants.  - Review and design the loan  selection process to include in the investment portfolio.  **Measure of success**:  - Design a prototype algorithm that would assess the default risk of the loans provided to an individual investor. |

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

|  |
| --- |
| **Population**: utilizing the accepted datasets given that investors would only fund approved loan applications. The initial screening/filtering process rests with Lending Club.  **Timeframe**: Even though the accepted dataset contains information from 2007 – 4Q2018, we would like to use the information starting from 2012 as LendingClub significantly ramped up its origination activities after 2011. Also, the dataset reflects the changing strategies of the company more closely.  **Target variable**: Feature that are available to investors at the time of investment. EDA in process.  Questions to answer with business and risk:  - Timing of default?  - Classification of delinquencies |

**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…

|  |
| --- |
| **Dataset 1: Accepted applications**  - approved data provided by Lending Club  - loan application amount, etc.  - 10 years of data → ~2.7mn instances with 151 features / columns. However, will be using data from 2012 onward.  - features include: request id, purpose loan amount, term, state, FICO score, … |

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

|  |
| --- |
| **Data preparation**  - Data munging and data scrubbing => prepare data to correct dtypes.  - Review and fill nulls  - Review feature relevancy  - Feature engineering => replace some features (e.g. use average FICO score instead of using both High and Low features)  **Target variable**  - Examine delinquency duration information  **Tools**  - Unzip dataset using gzip  - Data preparation in pandas  - data visualization packages in pandas and R  - Python machine learning package (Random Forest, Boosting, KNN, SVM, etc.)  **Analysis**  - EDA: statistical analyses => initial insights to share with stakeholders  - Prediction model: develop baseline models |

**4. Project plan**

**Kick off**

Project proposal and timeline

**Milestone 1**

- Validation assumptions/ choices on data

- Share initial insights

**Milestone 2**

- Share final results

- New decision process proposal

- Final presentation draft

**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* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data extraction (ETL) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Explor. data analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalize results |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Presentation prep |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Buffer |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 2** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Delivery** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |