**Statement of Work - SOW**

**Credit Risk Prediction**

**Prepared By:**

**Muhammad Imran**

**Capstone –AIDI 1003**

**Fall 2020**

**Executive Summary**

Credit Score is a common risk control operation in banking and financial industry. Financial institutions use various data sources to collect information about customer before approval of credit. The objective is to use historical data of applicants to predict whether or not an applicant will be able to repay a loan. Additionally, this is to make sure that applicant capable of repayment is not reject and consequently, avoiding loss to bank as well as customer.

**Rationale Statement**

Build machine learning models, in order make real-time lending decisions tailored to the individual customer. Applying supervised learning and classification algorithms to build models which provide better insights, accurate decision and find interesting pattern in customer data. Improve decision making process by correctly predicting probability of default for landing approval.

**Data Requirements**

Application and previous history is required for model building. The labels are included in the training data and train a model to predict the labels from the features. Details of the data required are as follows

**Application Data {train test):**

Static data for all applications. One row represents one loan in our data sample.

**Bureau:**

All client's previous credits history from credit

**Transaction and Balances:**

Customer credit transaction details from various channels like POS, cash Loans, and Credit cards.

**Credit Card Balances:**

Credit card monthly balances.**revious applications:**

All previous applications for Credit loans of clients who have loans in our sample.

**Repayment history:**

For the previously disbursed credits in Credit related to the loans in our sample.

\*\*note that all of the above have multiple features in file.

**Model/Architecture Approach**

**Supervised classification models:**

Credit risk is a traditional classification problem to find probability of default. In order to find the best result, various supervised classification models will be used and compare for best possible solution.

Programming language and Tools required are: Python notebook, Colab and Anaconda – Jupyter notebook.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Project Task - Sprint (1 week)** | **Start Date** | **End Date** | | Collect information & Analyze Dataset | 01-10-20 | 08-10-20 | | Clean Dataset | 08-10-20 | 15-10-20 | | Develop Data Preprocessing Pipeline | 15-10-20 | 22-10-20 | | Prototype algorithm | 22-10-20 | 29-10-20 | | Develop model(s)/architecture | 29-10-20 | 05-11-20 | | Train model(s)/architecture | 05-11-20 | 12-11-20 | | Test model(s)/architecture | 12-11-20 | 19-11-20 | | Evaluate model(s)/architecture | 19-11-20 | 26-11-20 | | Refine model(s)/architecture & Develop report | 03-12-20 | 10-12-20 | | Develop scorecard & Develop dashboard | 10-12-20 | 15-12-20 | |

\*\* Note that every sprint is reduced to 1 week as per capstone weekly class schedule.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Detailed Activities and Deliverable** |  |  | | **Collect information & Analyze Dataset** | 01-10-20 | 08-10-20 | | create labeling documentation Validate quality of data | **Outcome : Data with required features** | | |  |  |  | | **Clean Dataset** | 08-10-20 | 15-10-20 | | Exploratory Data Analysis | **Outcome : Quality Data** | | | Examine the Distribution of the Target Column | | Examine Missing Values, Column Types, Anomalies | |  |  |  | | **Develop Data Preprocessing Pipeline** | 15-10-20 | 22-10-20 | | preprocessing and modeling steps | **Outcome: Data Processing Pipeline** | | |  |  |  | | **Prototype algorithm** | 22-10-20 | 29-10-20 | | Build model prototype with represented data to interpret model | **Outcome: Prototype** | | |  |  |  | | **Develop model(s)/architecture** | 29-10-20 | 05-11-20 | | Establish baselines for model performance |  |  | | Start with a simple model using initial data pipeline | **Outcome : Baseline Model** | | | Overfit simple model to training data | | Try parallel ideas during early stages |  |  | |  |  |  | | **Train model(s)/architecture** | 05-11-20 | 12-11-20 | | Train model with clean data - low bias and variance | **Outcome : Trained Model** | | |  |  |  | | **Test model(s)/architecture** | 12-11-20 | 19-11-20 | | Write tests for:Input data pipeline |  |  | | Model inference functionality | **Outcome: Test Model with test Data** | | | Model inference performance on validation data |  |  | |  |  |  | | **Evaluate model(s)/architecture** | 19-11-20 | 26-11-20 | | ensure that this metric drives desirable downstream user behavior | **Outcome : Validate expected behavior** | | |  |  |  | | **Refine model(s)/architecture & Develop report** | 03-12-20 | 10-12-20 | | refine models - feature selection / additional learning layer | **Outcome: Improved Model** | | |  |  |  | | **Develop scorecard & Develop dashboard** | 10-12-20 | 15-12-20 | | build dash bord for visualization | **Outcome: Dash Board - story telling** | | |  |  |  | |

References:

<https://www.kaggle.com/rikdifos/credit-card-approval-prediction/discussion/119320>

<https://www.kaggle.com/c/home-credit-default-risk/data>

<http://archive.ics.uci.edu/ml/datasets/credit+approval>

[http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#](http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)

<https://www.kaggle.com/uciml/german-credit>