**Loan Defaulter Prediction Using Industry Segmentation and Sentiment Analysis**

**Overview**

This project contains code for predicting loan defaulters using various machine learning and deep learning techniques. It includes preprocessing steps, feature selection, model building, hyperparameter tuning, and evaluation. The main objective of this project is to predict loan defaulters using data from Lending Club, focusing on industry segmentation and sentiment analysis.

**Project Structure**

* Data: The data used for the analysis is the Lending Club data
* Notebooks:
  + Lending\_Club\_EDA.ipynb: Contains the Exploratory Data Analysis of the Lending Club dataset and merging of sentiment scores.
  + Feature\_set1.ipynb: Contains model building, and preprocessing steps for the data after removing variables with more than 0.85 correlation.
  + Feature\_set2.ipynb: Contains model building, and preprocessing steps for the data after removing variables with more than 0.6 correlation.
  + Feature\_set3.ipynb: Contains model building, and preprocessing steps for the data after removing variables with more than 0.6 correlation and collection fee variable.

**Dependencies /Libraries to Install:**

* imbalanced-learn==0.9.0
* numpy==1.22.2
* pandas==1.4.3
* scikit-learn==1.0
* seaborn==0.11.2
* tensorflow==2.8.0
* xgboost==1.5.0
* matplotlib==3.5.1
* hvplot== 0.9.2

**How to Run**

* Clone or download all the notebooks shared and the data file and place it all in the same directory OR download only Lending\_Club\_EDA.ipynb, Feature\_set2.ipynb(among the feature sets).
* **Data Preprocessing**: accepted\_2007\_to\_2018Q4.csv the lending club data is used to do the necessary pre-processing steps that are present in the “Lending\_Club\_EDA.ipynb” file where the missing values and outliers are handled, along with feature engineering to increase model performance and exploratory data analysis is conducted to understand the data better. sent\_updated.csv (Sentiment scores data) is utilized in this notebook for the industry sentiment scores which is merged in this notebook. The output after cleaning is saved in final\_data.csv.
* The Jupyter notebook “Feature\_set2.ipynb” contains the selected feature set code for which the steps are explained, the Feature\_set1.ipynb, Feature\_set3.ipynb contains the same process and steps.
* **Model Training**: The model training is done for the cleaned data which is final\_data\_v3.csv. The model is trained without hyperparameter tuning giving the basic model results, followed by hyperparameter tunning and a statistical test is performed for comparing the result of the model with and without the sentiment columns for each feature set.

1. **With sentiment**

* Basic model building – without hyperparameters
* Model building with grid search hyperparameters.
* Keras sequential model with three dense layers and 100 epochs trained.

1. **Without sentiment**

* Basic model building – without hyperparameters
* Model building with grid search hyperparameters.
* Keras sequential model with three dense layers and 100 epochs trained.

1. **Statistical validation**

There are 3 feature sets, and the respective csv data is utilized for the same and the steps remain the same.

* + **Feature set 1**: Removed all variables with correlation > 0.85(final\_data\_v2.csv)
  + **Feature set 2**: Removed all variables with correlation > 0.6(final\_data\_v3.csv)
  + **Feature set 3**: Removed all variables with correlation > 0.6 and collection fee variable. (final\_data\_v3.csv)
* **Model Evaluation:** The models are evaluated using the F1 score and Roc-Auc score. A t-test is performed to conclude if the addition of the sentiment and industry column has enhanced the ability of the model to predict loan defaulters.

**Results:**

The results are present at the end of the corresponding Jupyter notebooks. From the results, we can conclude that the addition of Sentiment and Industry columns has contributed to an increase in model accuracy for the prediction of loan defaulters. Additionally, the sentiment variables also appear as one of the top features for loan default prediction.

**Contact:**

**Name: Hemasri Appavu Krishnaraju**

**Email id:** [**hemasri.krishnaraju@ontariotechu.net**](mailto:hemasri.krishnaraju@ontariotechu.net)

**Mobile : +19052436716**