MODEL FOR INTEREST RATE PREDICTION

(GAUTAM GOWRISHANKAR)

# Project Report

## 1.1 SUMMARY DESCRIPTION

The goal of this project is to build a machine learning model that can predict the most appropriate interest rate for a given set of parameters provided by the Lending Club transactional loan data.

## 1.2 BACKGROUND

Lending Club is a peer to peer lending company, in which investors provide funds for potential borrowers and investors earn a profit depending on the risk they take (the borrowers credit score). Investors make money from interest. Based on the credit profile of each loan applicant, a grade IS assigned ranging from A1 to E5 with a corresponding interest rate. Investors make money from the interest rate and Lending Club receives a fee from both investors and borrowers, thereby establishing its business model

## 1.3 I/O Description

The file is a matrix of about 890 thousand observations and 75 variables. These files contain complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.

Kaggle provides a **‘loan.csv’** file and a dictionary for the various columns being used in the input file, it is in the form of .xlsx.

## 1.4 DESIGN EXPLANATION

Stage 1: Pre-processing

As part of the pre-processing stage, implemented Feature engineering. Feature engineering means building features for each label while filtering the data used for the feature based on the label’s cutoff time to make valid features. These features and labels are then passed to modeling where they will be used for training a machine learning algorithm.

Apart from Feature engineering, implemented Feature extraction using Random Forest library. Feature Selection is the process where we automatically or manually select those features which contribute most to our prediction variable or output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features.

Stage 2: Prediction Models

Neural Net, Random Forest and TPOT are the models that were implemented as part of prediction model creation. Achieved an accuracy of around 91% for Neural Net and 92% for

Random Forest. TPOT was able to produce as negative mse of 6.8%. With the help of matplotlib, visualized the predicted and actual values. Sickit was the main source library for

implementing all these models and for tpot, which is genetic algorithm based api, made use of a library called ‘TpotRegressor’

## 1.5 Instructions

Step1: Loan .csv file should be used as the input file and the code would produce cleaned.csv after manual pre-processing

Step 2: Cleaned.csv should be used as input for Pre-Processing notebook.

Step 3: Pre-Processing notebook will produce two CSV file namely, Features\_data.csv and Features\_Data\_final.csv

Step 4: Features\_Data\_final.csv should be used as the input file for executing Neural Net notebook

Step 5: Features\_Data\_final.csv should be used as the input file for executing Random Forest notebook

Step 6: pip install ‘tpot’ module before executing TPOT notebook.

Step 7: Cleaned.csv should be used as the input for TPOT notebook.

## 1.6 REFERENCES

[1] TDS, <https://towardsdatascience.com/feature-engineering-what-powers-machine-learning-93ab191bcc2d>

[2] Meium, <https://towardsdatascience.com/how-to-create-value-with-machine-learning-eb09585b332e>

[3] TPOT, <https://towardsdatascience.com/tpot-automated-machine-learning-in-python-4c063b3e5de9>

[4] NN, <https://developers.google.com/machine-learning/crash-course/introduction-to-neural-networks/anatomy>

[5] Feature Science, <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>