Assignment No 3 : Building a machine learning model using the Lending Club Dataset

Summary	Ten Use cases and motivations for building a machine learning model using the Lending Club Dataset
URL	
Category	Web
Environment	NA
Status	Version 1
Feedback Link	
Team No	7
Team Members	Maneendar Sorupaka (sorupaka.m@husky.neu.edu) Pratiksha Sawant (sawant.prat@husky.neu.edu) Rajsharavan Senthilvelan (senthilvelan.r@husky.neu.edu) Reema Mehta (mehta.r@husky.neu.edu)

Introduction

Our case study is build on peer-to-peer lending platform, which refers to the practice of lending money to individuals via online services that match anonymous lenders with the borrowers. Lenders can typically earn higher returns relative to savings and investment products offered by banking institutions.

Lending club established in 2008 is one of the largest peer-to-peer lender in the United States which issues loans between $1,000 and $40,000 for a duration of either 36 or 60 months. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015.

Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

In this case study we assume 10 fictional characters listed below who may be interested in working with Lending Club.

- *Rick* the Investor (Risk averse)
- *Tola* the Investor (Risk taker)
- *Taz* the Borrower (Good credit)
- *Pip* the Borrower (Bad credit)
- *Bipa* the Portfolio manager (Who lends to multiple people of different risk profiles)
- *Arc* the Arbitrager
- *Slick* the Data scientist
- *Irs* the Professor
- *Dat* the Data vendor who sells data and insights
- *Mar* the Regulator who regulates how the model can be used/not used

However before getting involved they need to be sure about the benefits and risks in business, so we analyze the Lending club dataset to build a prediction model to educate our client on the company and the insights drawn from the data to make intelligent decisions.

Task 1: Framing

Our client

Slick the Data scientist

Being a Data scientist Slick himself can study data of LendingClub and can analyze and predict what could be the possibility of low risk and what can provide returns.

"Optimized model with the most accurate interest rate prediction value"

He knows how to analyze data, but which model can be suitable based on the dataset and he can predict values.What slick is expecting from the prediction model is how accurately it will calculate interest rate value.

Statistics

Lending club shares its statistical information via graphs which gives wide information about its business and product performance.From 2010-2018 there is a significant increase in the total loan issuance for each quarter. The rise is evident in both the total loan amount issued as well as the number of loans issued.

As per the recent Borrower's Loan Purpose report about 68.81% is utilized to refinance existing loans(45.60%) or pay off credits(22.58%) and remaining 31.82% of loans are issued for car financing, home improvement, business etc.

Lending club widely issues loan throughout the States which could be determined from its loan issuance distribution represented geographically. More than $50M loan amount is issued in highly populated states like California, Texas, Washington, New York, Massachusetts whereas in states like Iowa, Wyoming it has issued maximum loan amount of $10M. A brief percentage count of loan issuance in States is given below:

Loan Issuance Range [$]	**States[Percentage]**
50+ M	40
25M - 50M	24
10M - 25M	16
0 - 10M	20

The investor account performance is determined by Adjusted Net Annualized Return of all the investor accounts on the Lending Club Platform that have invested in at least 100 Notes and that have not purchased or sold Notes on the Folio Investing Note Trading Platform. Investment returns can change over a period of time and different factors can influence the volatility of returns. Some of these factors are given below :

a. *Number of Notes in a portfolio*: Owning a small number of notes leads to more volatile returns. However, if the investment is distributed evenly across multiple notes corresponding to different Borrowers gives stable returns. This is called diversification.
b. *Concentration of investments*: If more than 50% of investment is concentrated to single loan and when the loan is fully performing the returns will be high however if the loan charges off the account value will decline substantially and returns may be low. Hence concentrating an investment may lead to high volatility in returns.
c. *Weighted Average Interest Rate*: More stable returns can be achieved from Notes with a lower interest rate because these loans are expected to have lower charge-offs. Notes with a higher rate of interest have more chances of declining.
d. *Weighted Average Note Age*:Having a higher or lower number of notes, will eventually going to decrease the returns associated with the accounts as these accounts could be expired, defaults or charged off.
e. *Composition of Portfolio*: Accounts with same average weighted rate interest can have different composition notes and hence have different returns.Example an account which holds Notes in B and C grades loan will have more stable returns as compared to the account which includes notes in all grades.
f. *Performance*: Accounts with a higher number of charges off will affect the overall performance of the portfolio.
g. *Vintage*:Returns can depend on the timeframe of investment. Market conditions can vary which will impact on the investment rates.

Average weighted interest rate:

Grades (Interest Rates)	A	B	C	D	E	F	G
36/60/All	5-10%	10-15%	10-17%	15-20%	15-25%	19-31%	20-30%

For grade A and B average weighted interest rate is stable in the given time period. While for grade C and D it is quite fluctuating. And for grade E, F and G it is gradually increasing over the years.

Total No. of loans issued:

The total number of loans issued for a period of 36 months for grade A and B is itself close to 65% for the tenure. Approximately 20% and 10% of the loan issued in grade C and D respectively. While grade E is having some number of loans issued, E and F are negligible.

For 60 months time periods, statics are different. Grade C is having the highest number of loans issued over the years. And grade B, D and E is showing a fairly good percentage. Grade E shows few loans issued while grade A and G have hardly loans issued.

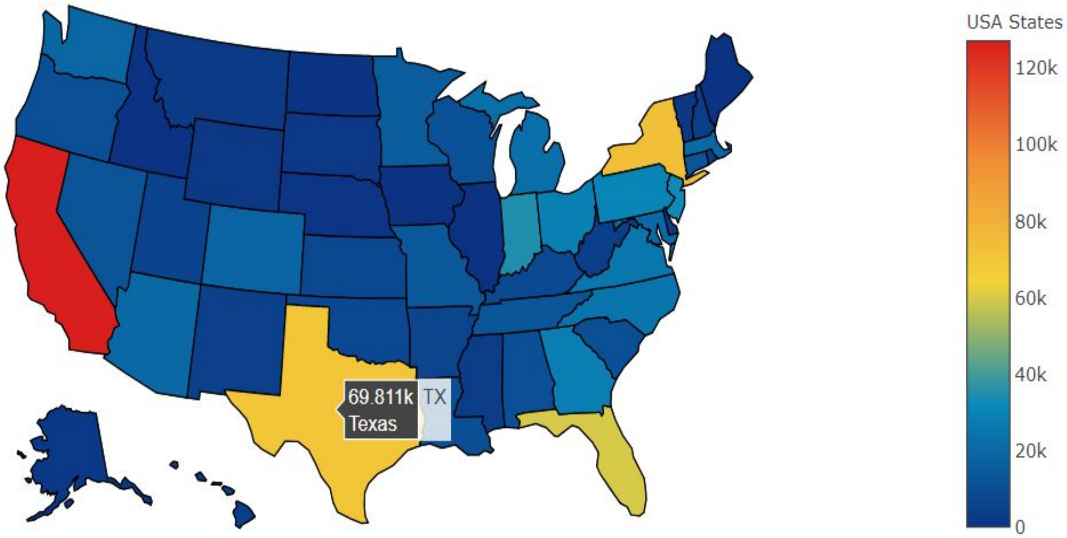
Loan Performance details:

Based on numbers we can depict that A is having low-interest rate but it will provide constant return with the least probability of getting the charge off. Grade B and C seems a suitable option with average interest and low risk of charge off. Higher interest rate will give higher returns but also lead to high chances of charge off.

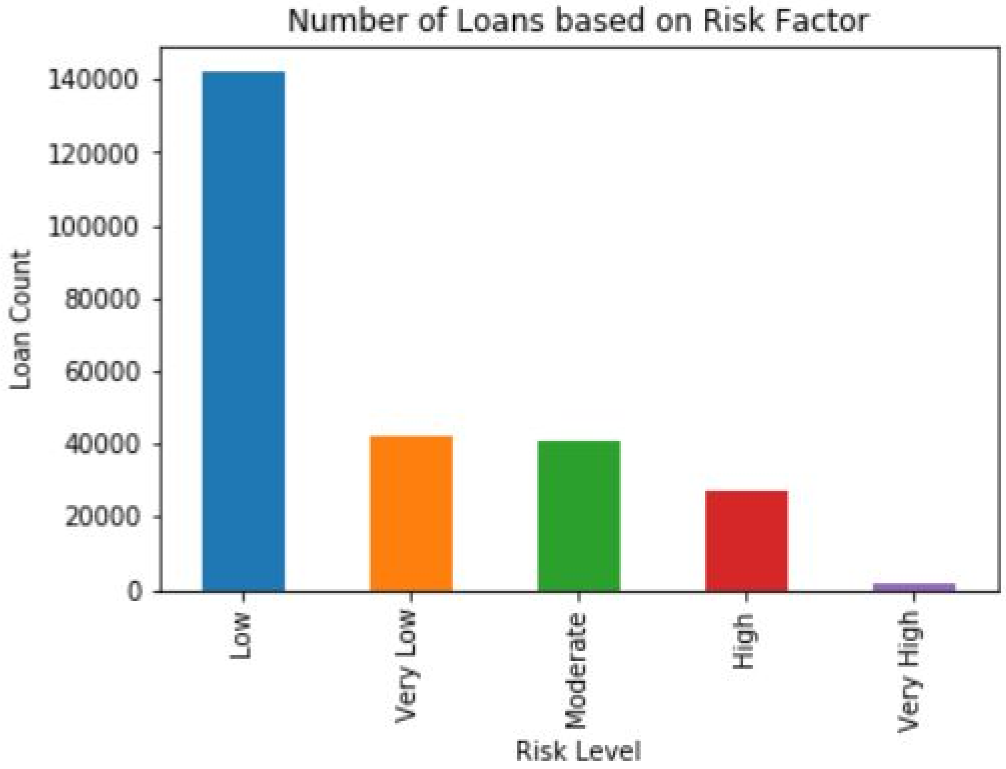
Over the period of time rates will get lower for the investment With delaying for payment in interest rate would higher the further amount before it will get charge off.

Insights

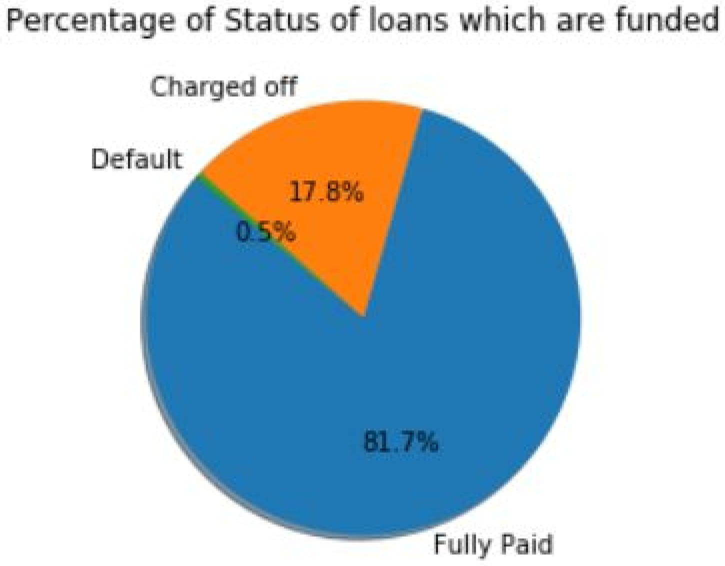
Based on the given data, we computed the number of loans processed through LC. As a geographical distribution, we found that California, Texas, New York are the states have a huge number of loan requests. Area, Population, Standard of living many factors can impact these requests.



Risk Factor is segregated based on which grade it belongs to, and counted the number of loans which is under, provided risk factor. Based on this data, we can see that most borrowers are going for low-risk loan option which is grade B and C, which will affordable and in range of their requirements.



Based on the data, showing that approximately 82% of loans are fully paid.Which will definitely a good number for an investor to invest in Lending Club



Task 2: Data Preparation

For data cleaning we eliminated some columns (open_acc_6m, verification_status_joint, emp_title, mths_since_last_record, out_prncp_inv etc) which were not required for our analysis.
We handled the missing values in the data by replacing null values using fillna method.
We replaced the null values by minimum values and using backfill technique. We also removed all the instances which represented loans which are not in status "Fully Paid" , "Default" and "Charged Off"

For extraction of features, from preprocessed data file we used both feature tools(Automated feature engineering) and manual feature engineering method. And from our experience, we observed both pros. and the cons. of each method.

Feature tools VS Manual feature engineering

Differentiation Factors	Feature tools	Manual feature engineering
Processing Time	It completely depends on the size of the dataset, increases with the huge dataset	It depends on which features you want to add for your model
Number of features	Gives all the possible permutation and combination for the entire dataset	Number of features will help to analyze a dataset with a different outcome

Accuracy	Each feature is fetched accurately	As working manually and testing your output will provide accurate data.
Specification	Could only achieve default specifications generated by feature tools eg: mean, mode, sum, min and max	Can add some extra specifications by manually analyzing the data such as deciding the ‘interest range’ for each loan amount

Task 3: Prediction

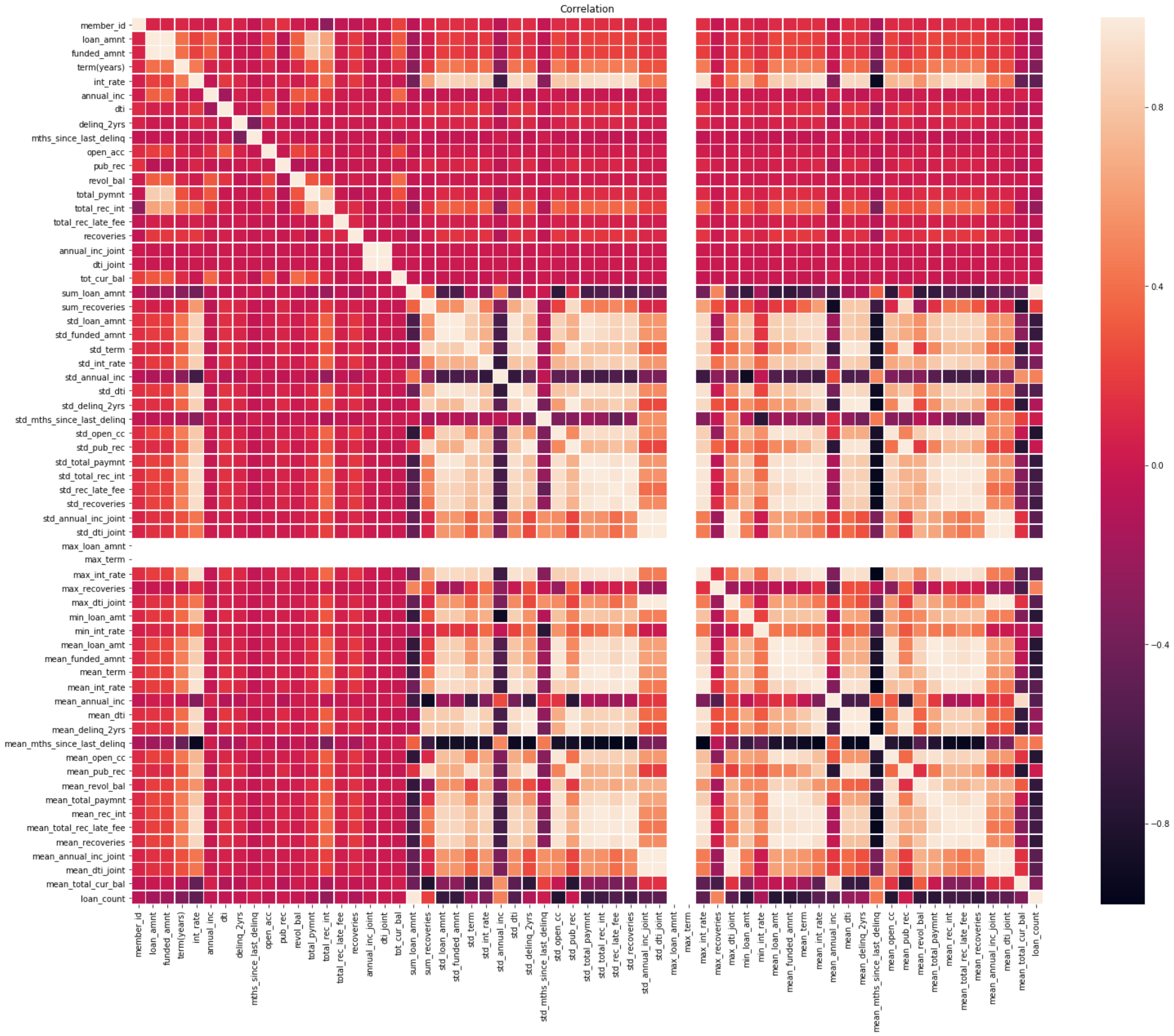
Regression Model:

Linear regression predicts a dependent variable (y) and a series of changing independent variables (x). It is assumed that there is approximately some relation between x and y which can be mathematically represented as $Y = mX + b + e$ where 'b' and 'm' are two unknown constants that represent intercept and slope and 'e' represents the errors.

Relation between each variable can be deduced by correlation. Correlation is an index that ranges from -1 to 1 which can be interpreted as follows:

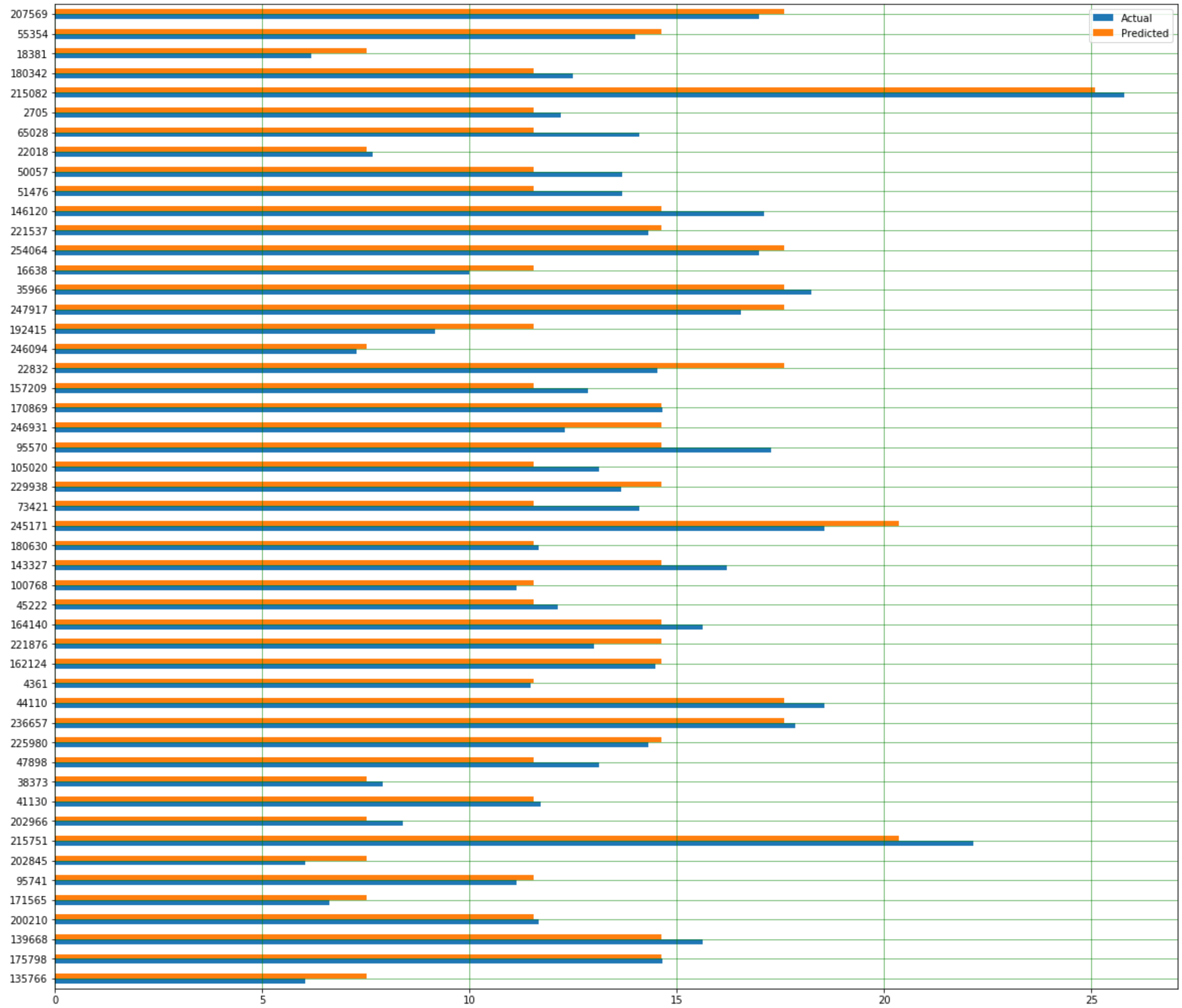
- When the value is zero there is no linear relationship between the variables
- When the value gets closer to +1 or -1 the relationship is stronger.
- When the value is 1 it indicates a linear relationship

To find out the correlation between variables we plot a heatmap as displayed below:



Further to implement the model we prepare training and testing datasets with *train_test_split()*, as we need to predict the interest rate so it is the dependent variable and rest are independent variables splitting the data into training and testing data.

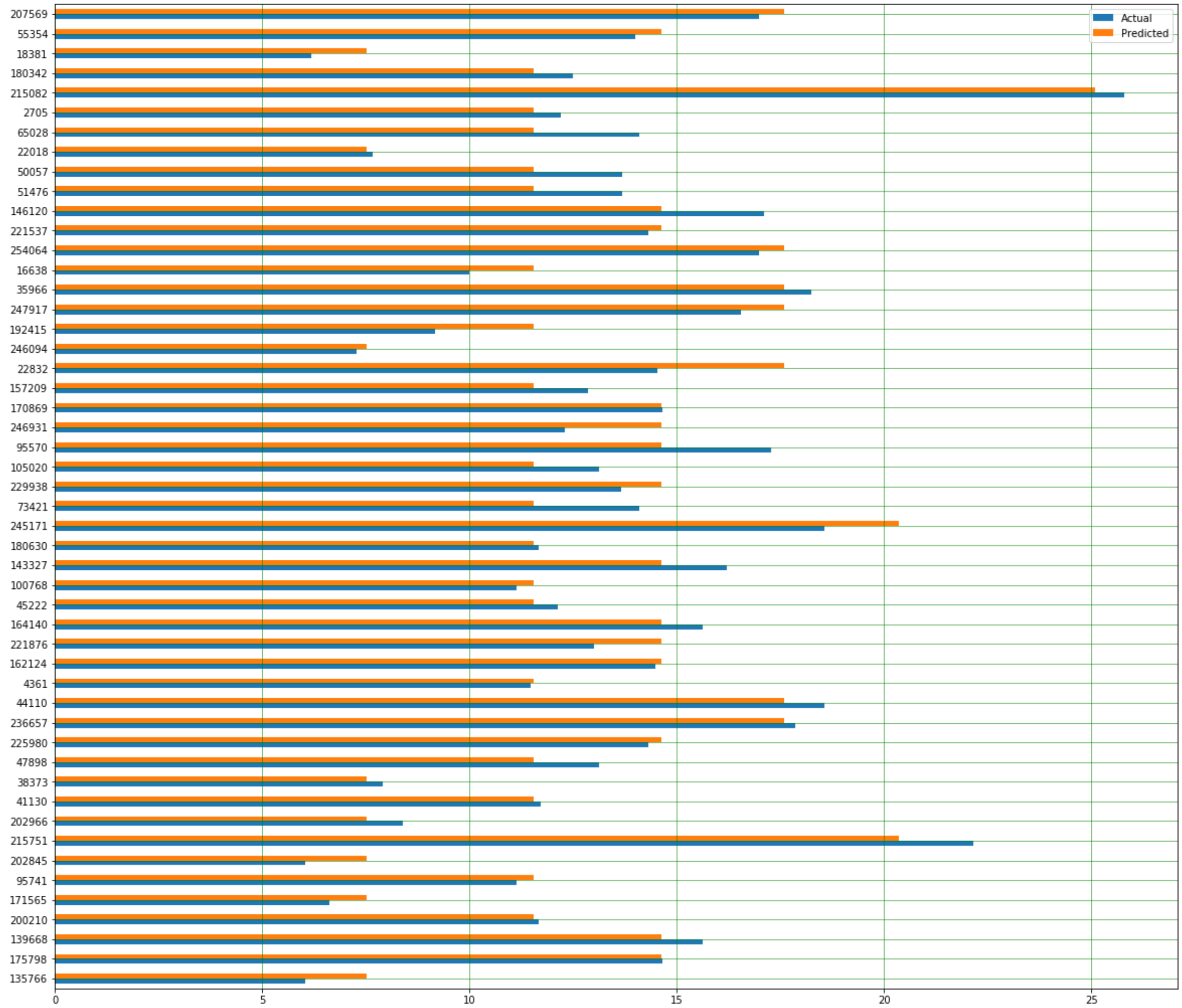
Then we train the model to predict the interest rate by fitting the model to the training dataset. It tries to find the best value of intercept and slope which results in the line that best fits the data. The prediction is visually represented using matplotlib as a horizontal bar for understanding the behavior



We calculate the MAPE value to evaluate the performance of the model. The MAPE for the linear regression model comes up to *8.436321371390914*.

5-Fold Cross-Validation

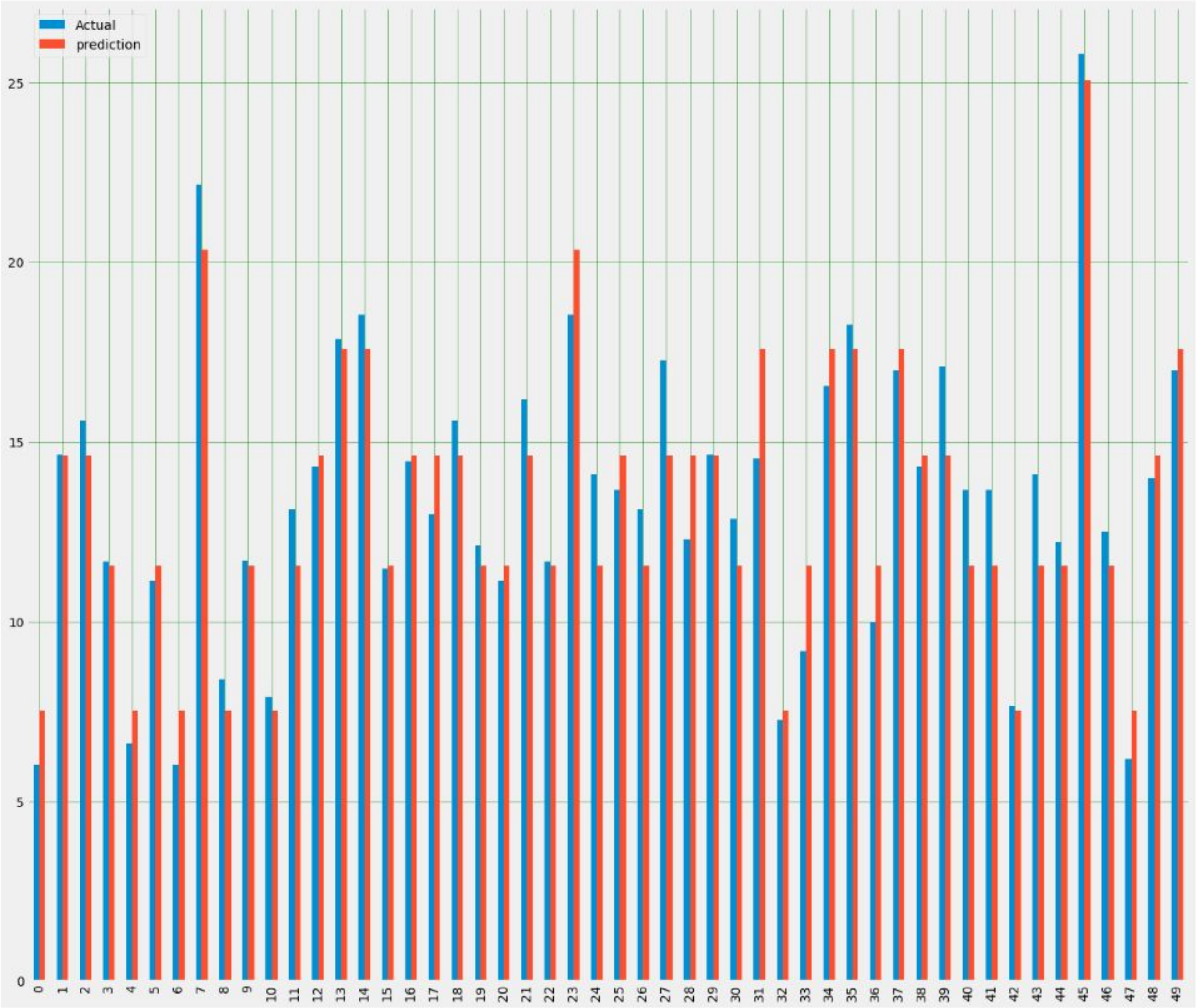
We perform 5-Fold cross-validation to get an accurate split for data for a more efficient model. In K-Folds Cross Validation we split our data into k different subsets (or folds). We use k-1 subsets to train our data and leave the last subset (or the last fold) as test data. We then average the model against each of the folds and then finalize our model. We split our data into 5 different subsets/folds to achieve prediction for an interest rate as below:



The MAPE value calculated after performing 5-Fold cross verification is *9.87527400681246*. Hence we can say that this step did not enhance the performance of the model.

Random Forest:

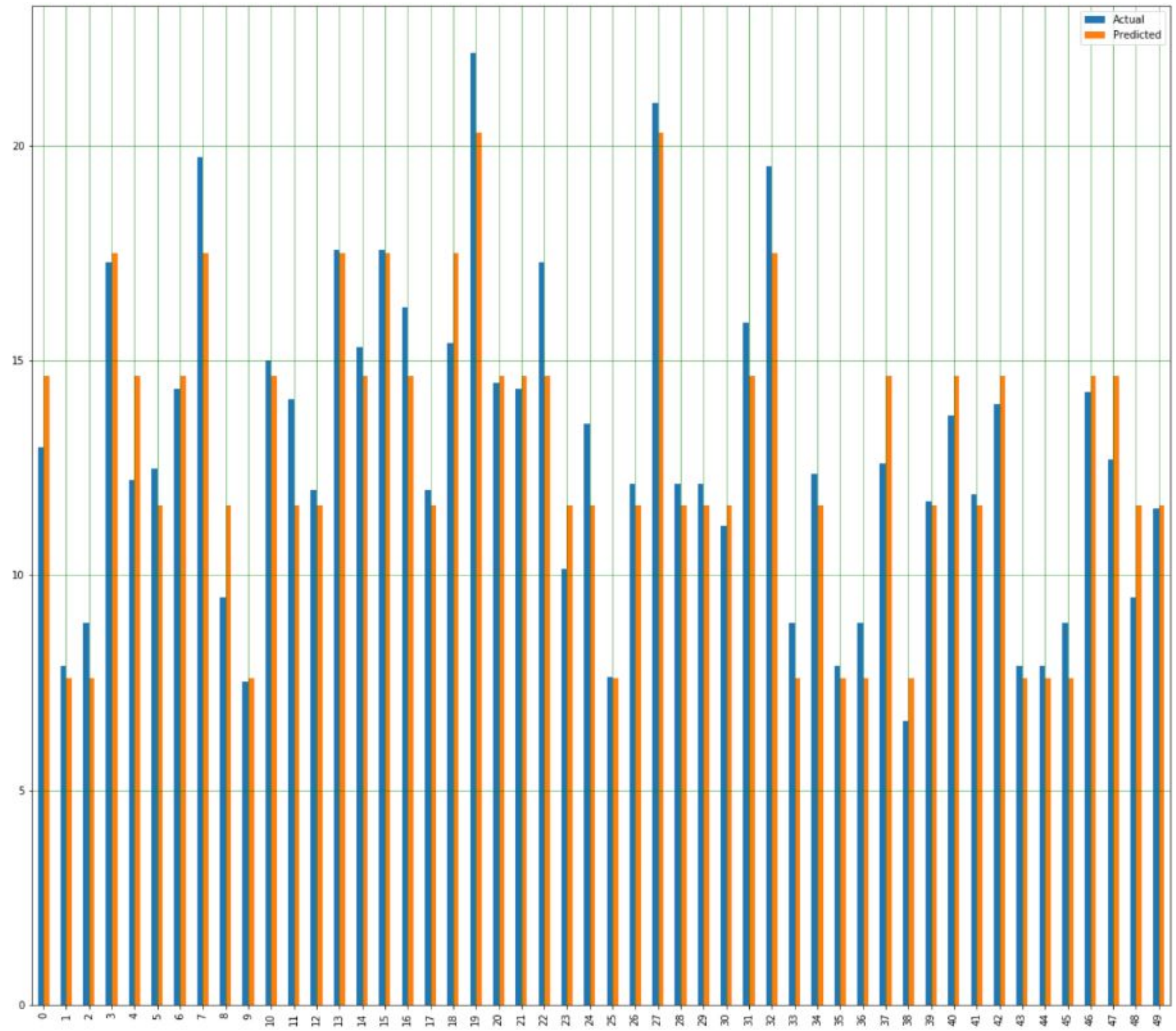
Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.



Neural Networks:

Neural networks are a set of algorithms, modeled loosely after the human brain, that is designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

Below graph is representing Actual vs predicted values.And this output has take after 5 cross-validation


Actual
Predicted
0
5
10
15
20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49

Task 3: Prediction Model Comparison

Based on the values we have received after performing 5 Cross-Validation on predictional models Regression, Random Forest and Neural Networks.

Factors	Linear Regression Model	Random Forest Model	Neural Networks
MAPE	9.875653344353633	**8.434945851378117**	8.480105313425364
MAE	1.1936371321843193	1.0582810809476917	1.0689406841087032
MSE	2.150445343304537	1.7447602861359501	1.7729289906207173
RMS	1.466439682804764	1.3208937452103973	1.3315137966317576
Accuracy	91.56%	**91.57 %**	91.52%

MAPE value will give mean of percentage error across the dataset.

lower value of MAPE indicates that error between actual and predicted output is smaller than the other model.

For our prediction, we got lower value of MAPE for Random forest and high accuracy value.

After performing **5 cross validation** on models, we observed that MAPE value is getting lower, but for our random forest, there was drastics change in MAPE value also accuracy increased by 17%

Task 4[a]: Hyper-parameter optimization

1. **Regression(Lasso, Ridge, and Elastic)**

Factors	**Lasso(L1)**	**Ridge(L2)**	**Elastic**
MAPE	8.451402539762718	8.442277175603637	8.459643245860294

MAE	1.0651620514930171	1.064397246552341	1.066328278442036
MSE	1.767377021712332	1.7674216170865387	1.7693346558727612

2. **Random Forest(No of trees, maximum depth)**

   Mean Absolute Error: 1.06 degrees
   MAPE value is: 8.434968237458582

3. **Neural Networks**

MAPE	8.419461005596112
MAE	1.0648073851380802
MSE	1.7898485296838647
RMS	1.337852207713492

Hyperparameter tuning is providing two important aspect.

First is the feature selection, where out of whole dataset, it will provide which are the most required features which can help us to predict most accurate interest rate.

And other one is help to reduce overfitting of model.

Task 4[b]: AutoML

Automatic ML tools are used to abstract all of the decision making that would have been done during an ML workflow. The selection of the model, its hyper-parameters, loss function and optimizing algorithm are all handled automatically. This reduces the ML practitioner’s task to data preparation and feeding into the model. Thereafter the automatic ML tools take care of everything else and return a highly optimized model after training on the availed data.

TPOT:
TPOT is a python Automated Machine Learning Tool that optimizes machine learning pipelines using genetic programming (inspired by Darwinian Process of Natural Solutions i.e. finding out the fittest possible solution for optimization). It automates the most tedious part of machine learning i.e data preparation, feature selection, feature engineering, model selection and validation, hyperparameter tuning and outputs the optimal code for you when it’s done. TPOT tool generates a CV score of *-6.846951416332602* for LC data. This is the mean squared error for our model. Automated machine learning has significantly improved on that score with a drastic reduction in the amount of development time

MAPE - 15.599448586333253

H2O.ai:
H2O's AutoML can be utilized for mechanizing the AI work process, which incorporates programmed preparing and tuning of numerous models inside a client indicated time-limit. Stacked Ensembles – one dependent on all recently prepared models, another on the best model of every family – will be consequently prepared on accumulations of individual models to create profoundly prescient outfit models which, by and large, will be the best performing models in the AutoML Leaderboard.

MAPE - 0.21457457356189942

AutoSKLearn:
One of the intuitive automatic machine learning toolkits is known as Auto-SKlearn. It is based on the highly popular Scikit-Learn Python library and is written as a wrapper above the original library. It makes calls to underlying Scikit-Learn functions and passes parameters to them automatically. Auto-sklearn acts as a replacement for a scikit estimator and frees the ML practitioner from model selection and tuning. As such it cuts down tremendously on the ML life cycle and allows quick iterations to explore a dataset and potential ways to model it.

MAPE - 9.131635495565906

Manual vs Auto ML

Factor	Linear Regression Model	Random Forest Model	Neural Networks

MAPE	9.875653344353633	**8.434945851378117**	8.480105313425364
Factor	**TPOT**	**H2O.ai**	**AutoSKLearn**
MAPE	15.599448586333253	**0.21457457356189942**	9.131635495565906

Task 5: Analysis

Prediction Model – Random Forest

We performed feature engineering on preprocessed data of Lending club using feature tools. We calculated Mean absolute percentage error(MAPE) value for training and testing data set using Regression, Random Forest and Neural network model. And from the values, we concluded that the Random Forest is the most suitable model for predicting interest rate from the dataset which we provided.

Factor	Linear Regression Model	Random Forest Model	Neural Networks
MAPE	9.875653344353633	**8.434945851378117**	8.480105313425364

1. In this model we have performed feature reduction and limiting the number of features for prediction.
2. Performed Random search with cross validation and hyperparameter tuning and evaluated output with baseline model and which gave significant change in accuracy.
3. We performed grid search which moved accuracy with minor number, so optimized it to the best for output.

4. At the end we performed grid search with cross validation, which provided the most optimizing values, which will be helpful for prediction.

Random Forest model is providing 91.57% accuracy, which can help to our model to get closest predicted values to actual values.

Optimization - Without the variable importances, the value of MAPE is 16.59 and after optimizing by providing the most important variables and Hyperparameter tuning number of tree and depth, value came down to 8.43.It shows we can optimize output by applying different and required hyperparameter tuning.

Performance - It is reducing features and giving the %of most important features.Only relevant features can be taken for prediction.

For Data Scientist, this could be the best model which provides the best output for prediction of interest rate for both borrowers and Investors.By passing features to narrow down those features in particular manner.And then providing result with accuracy percentage of 91.57%.

Languages and Tools Used

Language	Python 3.7
Libraries/Tools	numpy,pandas, featuretools, seaborn, sklearn, TPOT, AutSKLearn, H2O
Visualization	Matplotlib, plotly

Citations

https://www.liebertpub.com/doi/full/10.1089/big.2018.0092

https://www.investopedia.com

https://towardsdatascience.com/automated-feature-engineering-in-python-99baf11cc219

https://docs.featuretools.com/automated_feature_engineering/primitives.html

https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f

https://towardsdatascience.com/linear-regression-using-python-ce21aa90ade6

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://github.com/EpistasisLab/tpot

https://www.datacamp.com/community/tutorials/tpot-machine-learning-python

https://stackabuse.com/using-plotly-library-for-interactive-data-visualization-in-python/

https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

https://medium.com/@jayeshbahire/lasso-ridge-and-elastic-net-regularization-4807897cb722

https://github.com/agu3rra/NeuralNetwork-RegressionExample/blob/master/Tutorial.ipynb

https://github.com/h2oai/h2o-tutorials/blob/master/h2o-world-2017/automl/Python/automl_regression_powerplant_output.ipynb

https://github.com/h2oai/h2o-tutorials/blob/master/tutorials/custom_metric_func/CustomMetricFuncRegression.ipynb

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html