

# Lending Club Loan Prediction

## Authors

Bhavya Haridas - Northeastern University  
Harshitha S Gadadhar - Northeastern University  
Saumil Shah - Northeastern University  
Nikita Gawde - Northeastern University

---

## Introduction

Lending Club is a peer to peer lending company based in the United States, where investors provide funds for potential borrowers to earn a profit depending on the risk they take. Lending Club provides acts as the platform between investors and borrowers.

We are a group of analysts guiding our client Bipa in designing a portfolio that maximizes her return.

The Lending Club loan data available is used to predict the interest rate she can earn for her profiles.

---

## Task 1 - ML Goals and EDA

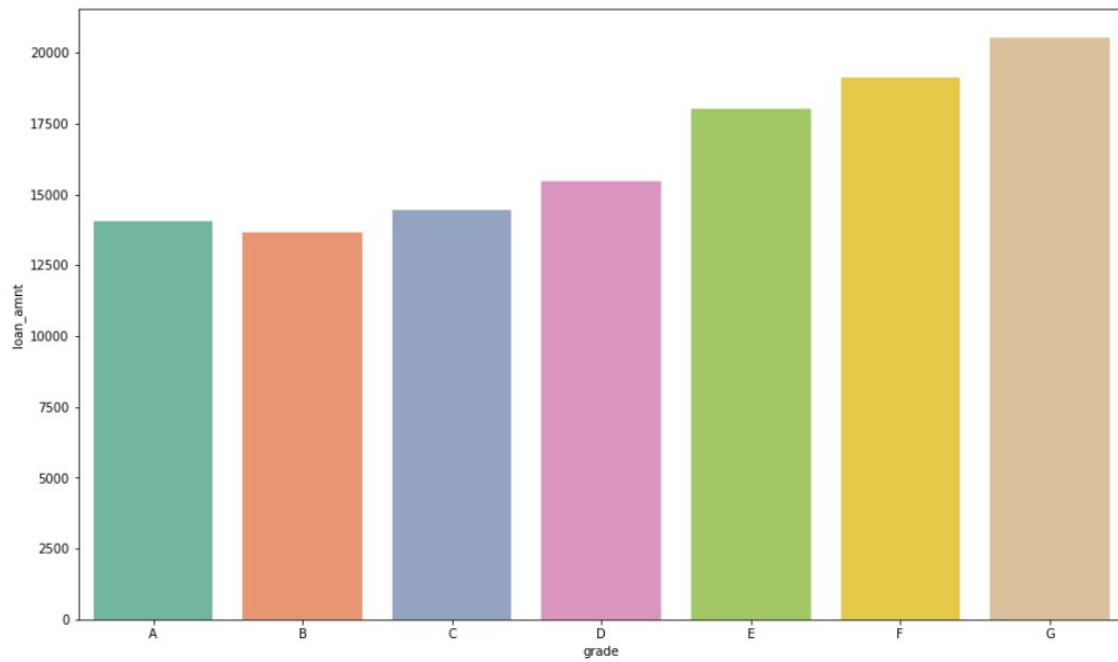
### Client - Bipa, the portfolio manager

- Predict interest rates of each of her profile
- Invest in various risk profiles
- Manage Risk and have a long term profit
- Analyze options for partial loan investments in building a portfolio.
- Find the right portfolio size to maximize profit

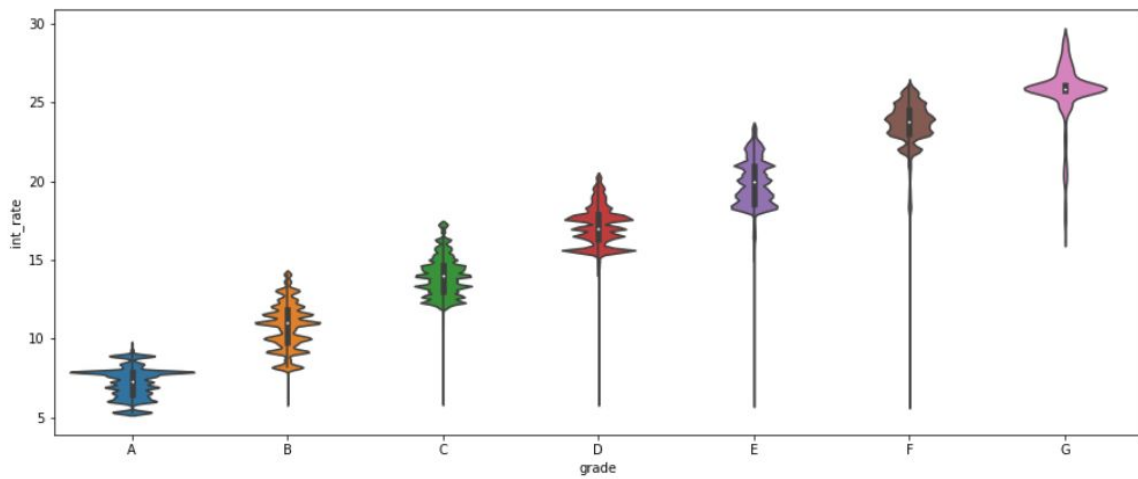
## Data Exploration

- Dropped the most trivial columns - 'url', 'desc', 'policy\_code', 'emp\_title', 'sub\_grade', 'verification\_status\_joint', 'member\_id', 'title', 'zip\_code', 'initial\_list\_status', 'pymnt\_plan'

Loan Amount v/s grade



Interest Rate v/s Grade



Interest Rate increases with increase in risk

## Percentage of loan fully paid off in each Grade

% of loans that are current or fully paid by each group of grade

A: 97.51%

B: 94.69%

C: 92.21%

D: 88.75%

E: 86.22%

F: 81.02%

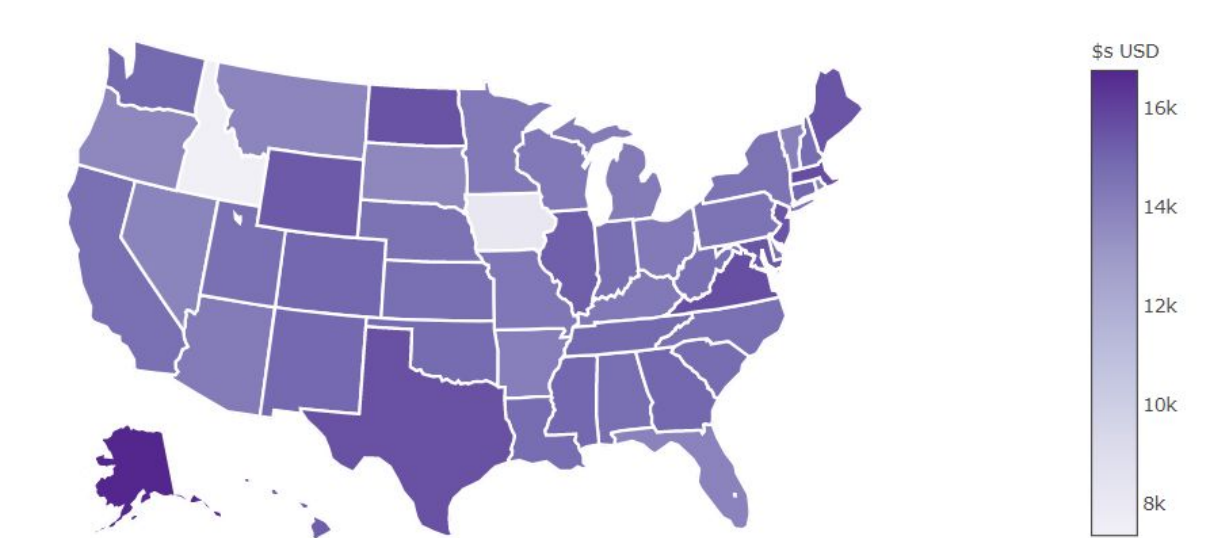
G: 77.24%

## Number of loans with particular status in each grade

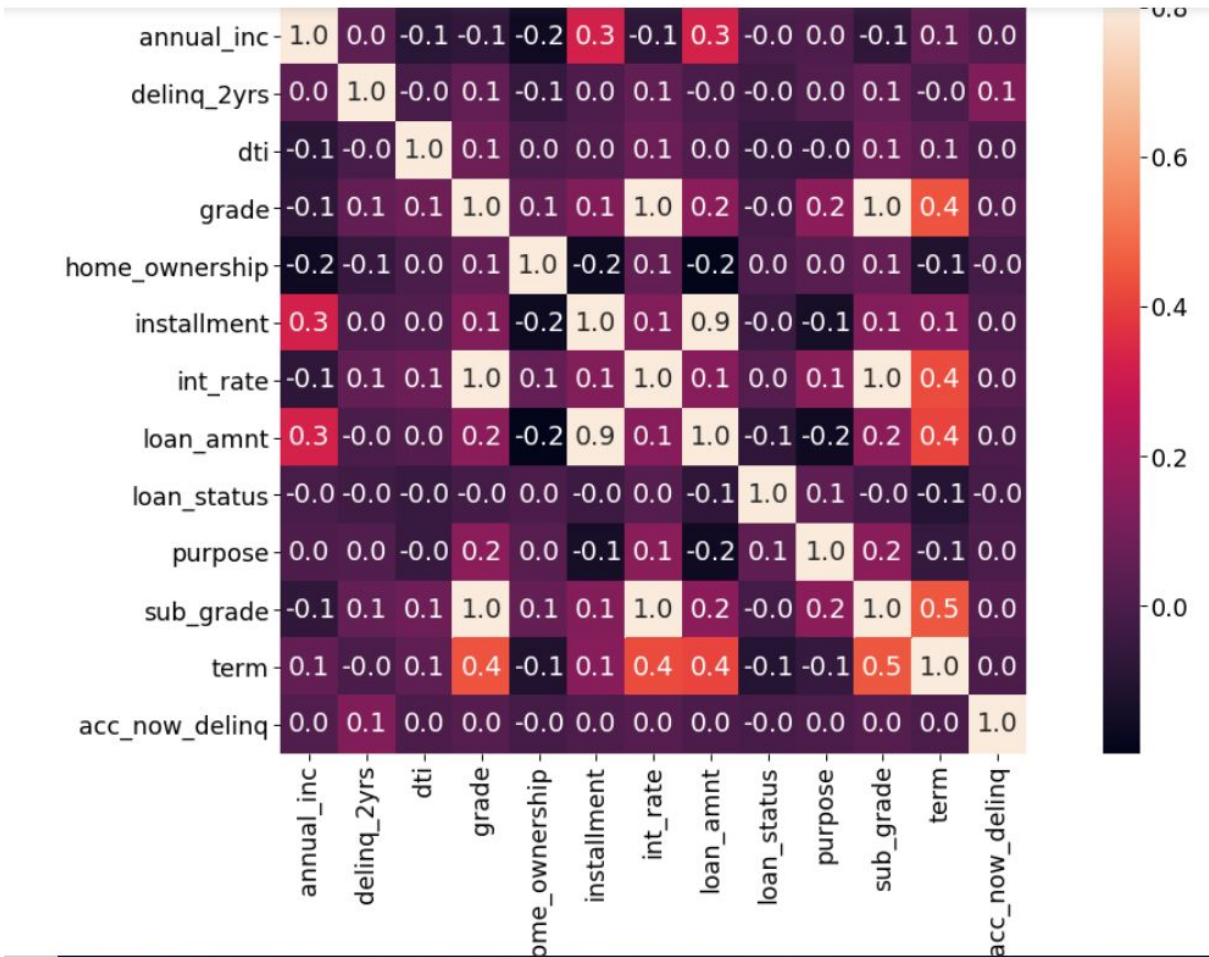
grade	A	B	C	D	E	F	G
loan_status							
Charged Off	2617	9519	12642	10486	6258	2934	792
Current	103322	171735	171175	91984	47061	13589	2913
Default	47	198	360	312	201	79	22
Does not meet the credit policy. Status:Charged Off	8	85	148	197	158	93	72
Does not meet the credit policy. Status:Fully Paid	90	269	481	494	378	154	122
Fully Paid	39679	66546	52678	30020	12928	4726	1146
In Grace Period	365	1240	1887	1405	908	354	94
Issued	1448	2529	2472	1185	593	194	39
Late (16-30 days)	134	410	678	569	368	155	43
Late (31-120 days)	492	2004	3339	2890	1852	768	246

Grade is not entirely dependent on status of loan

Loan Issued by state



Correlation



- sub\_grade and term seem have a high correlation.
- loan\_status is not correlated with most of the factors

### Home Ownership v/s Status

home_ownership	ANY	MORTGAGE	NONE	OTHER	OWN	RENT
loan_status						
Charged Off	0	19878	7	27	4025	21311
Current	2	303764	2	3	62041	235967
Default	0	498	0	0	110	611
Does not meet the credit policy. Status:Charged Off	0	348	1	11	49	352
Does not meet the credit policy. Status:Fully Paid	0	908	4	27	138	911
Fully Paid	1	104966	36	114	17960	84646
In Grace Period	0	2855	0	0	637	2761
Issued	0	4220	0	0	1038	3202
Late (16-30 days)	0	1101	0	0	260	996
Late (31-120 days)	0	5019	0	0	1212	5360

- Highest fully paid loan borrowers are on mortgage
- Highest defaulted and charged off borrowers are on rent

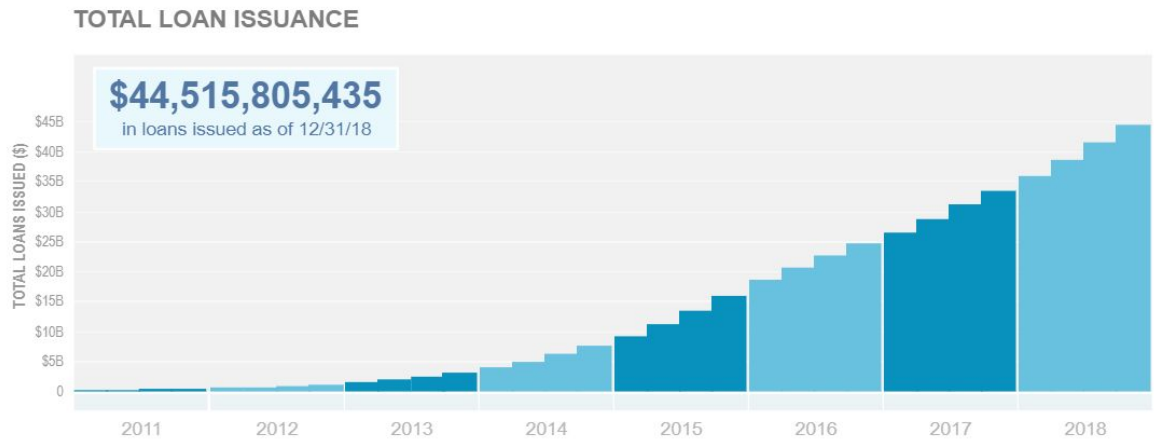
### Application type v/s Interest Rate



- Higher Interest rate is offered for joint applications

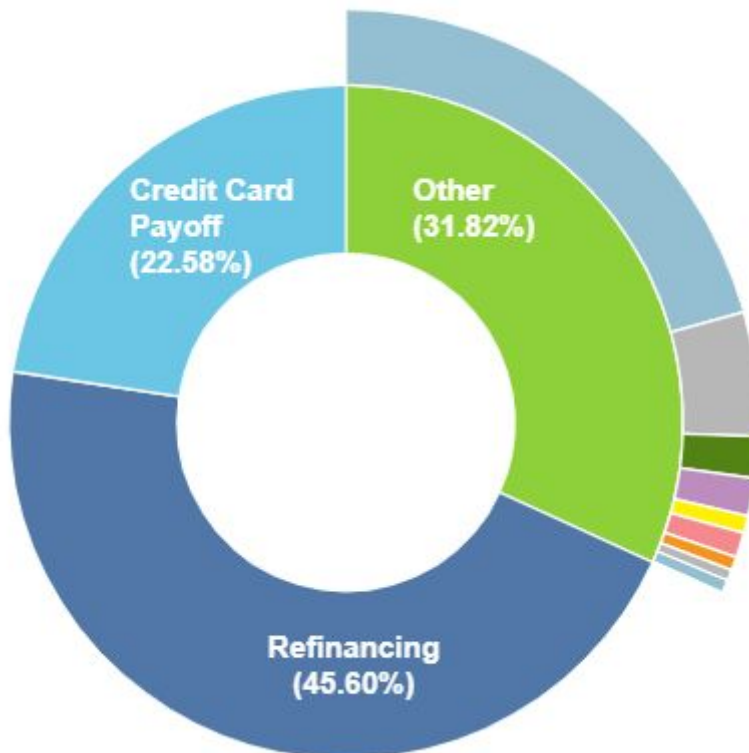
## Key Take-away from Lending Club Charts

Number of loans over the years



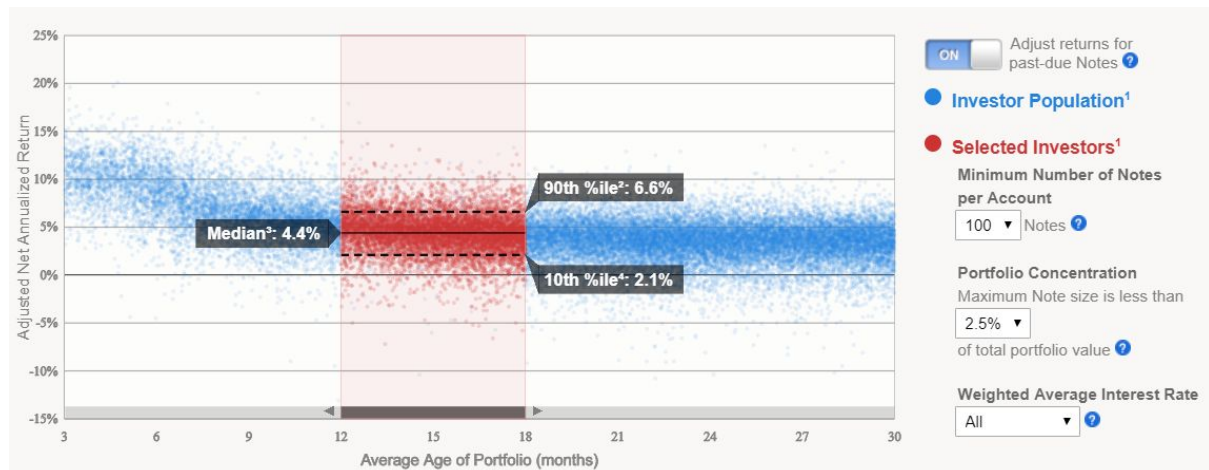
- Assures the client of the growing popularity of Lending Club

Purpose



- Majority in refinance - assures client of lower interest rate loans to borrow and re-invest in other domains

## Investor Account Returns by Average Age of Portfolio



- Lending club investors with diverse accounts usually have better returns when compared to concentrated holdings.
- Diversification increases when additional notes are purchased with respect to different borrower loans.

Accounts with more than 100 notes of only grades A to E have likely to have positive returns. For example, If Bipra needs to invest \$2000 in Lending club , instead of investing the entire amount in one borrower , she can invest varied amounts in different borrowers like \$20 in 100 different accounts.

## Task 2 - Data Preprocessing and Feature Engineering

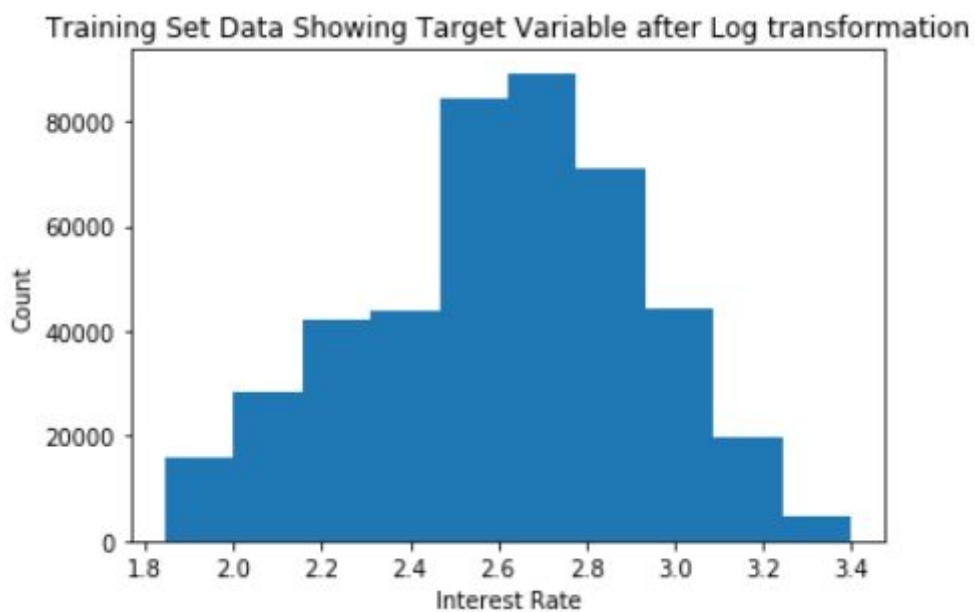
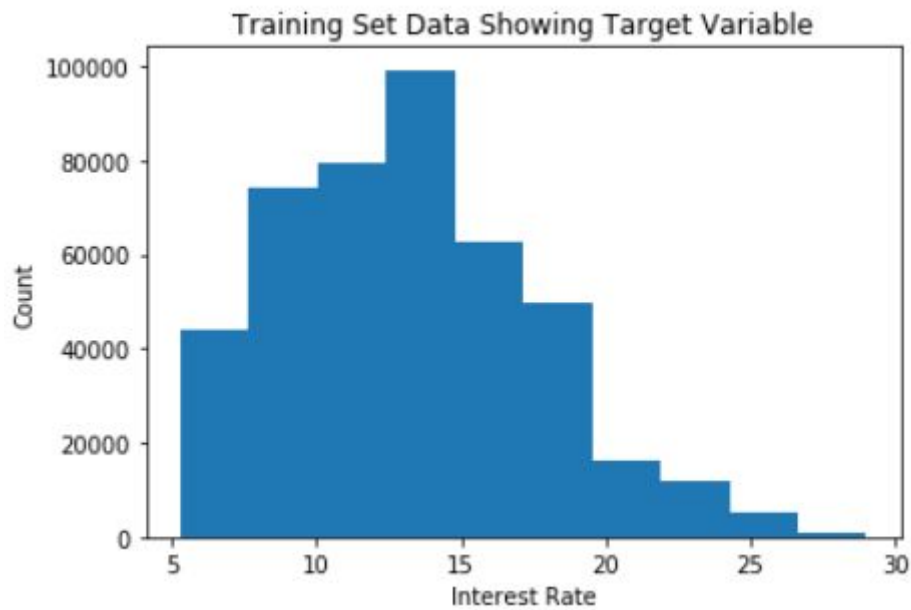
### Data Cleaning

- Read data, drop some of the most irrelevant fields and clean data to get rid of unwanted characters

loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_s
5000.0	5000.0	4975.0	36.0	10.65	162.87	2	10.0	6	24000.0	3	22	
2500.0	2500.0	2500.0	60.0	15.27	59.83	3	10.0	6	30000.0	2	22	
2400.0	2400.0	2400.0	36.0	15.96	84.33	3	10.0	6	12252.0	1	22	
10000.0	10000.0	10000.0	36.0	13.49	339.31	3	10.0	6	49200.0	2	22	
3000.0	3000.0	3000.0	60.0	12.69	67.79	2	1.0	6	80000.0	2	22	



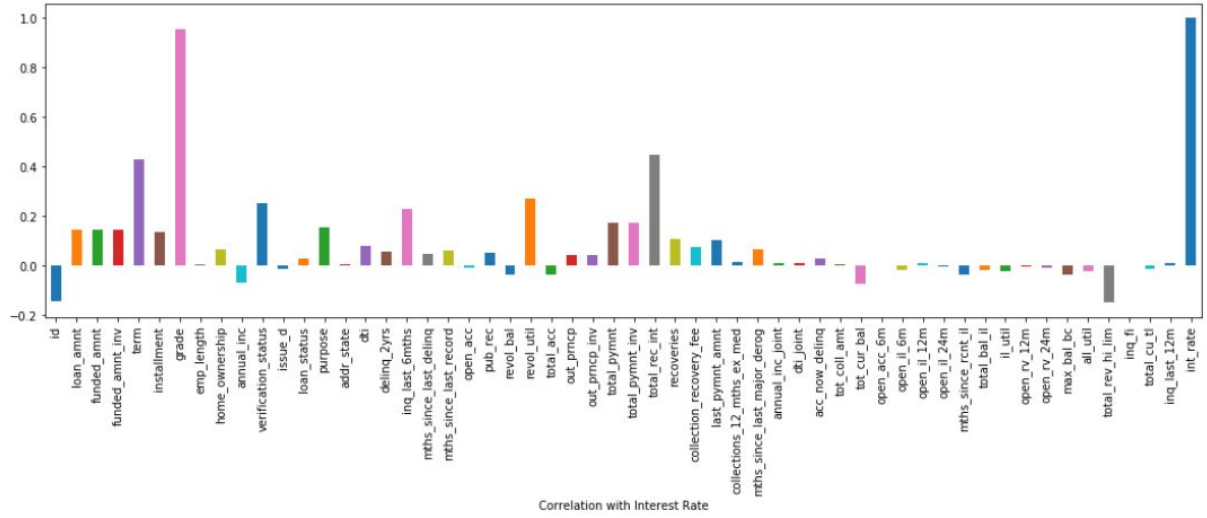
## Variable Selection



## Feature Correlation

Correlation with Interest Rate





## Top Features out of Manual Feature Engineering

int_rate	1.000000
grade	0.954166
total_rec_int	0.446664
term	0.428699
revol_util	0.268713
verification_status	0.253051
inq_last_6mths	0.227664
total_pymnt_inv	0.172735
total_pymnt	0.171807
purpose	0.151558
funded_amnt_inv	0.145904
funded_amnt	0.145881
loan_amnt	0.145713
installment	0.133580
recoveries	0.106086
last_pymnt_amnt	0.101673
dti	0.079648
collection_recovery_fee	0.074119
mths_since_last_major_derog	0.064476
home_ownership	0.064383
mths_since_last_record	0.062098
delinq_2yrs	0.056741
pub_rec	0.052501
mths_since_last_delinq	0.046904
out_prncp	0.041384
out_prncp_inv	0.041242
loan_status	0.028100
acc_now_delinq	0.026760
collections_12_mths_ex_med	0.014739
dti_joint	0.011105
inq_last_12m	0.009413
annual_inc_joint	0.007502
open_il_12m	0.006787
addr_state	0.005920
tot_coll_amt	0.004326
emp_length	0.003544
open_acc_6m	0.001783
inq_fi	-0.000722
..	- - - - -

## Feature Engineering with FeatureTools

- Generated 332 features

```
Index(['SUM(dat.loan_amnt)', 'SUM(dat.funded_amnt)',
      'SUM(dat.funded_amnt_inv)', 'SUM(dat.term)', 'SUM(dat.int_rate)',
      'SUM(dat.installment)', 'SUM(dat.grade)', 'SUM(dat.emp_length)',
      'SUM(dat.home_ownership)', 'SUM(dat.annual_inc)',
      ...
      'MEAN(dat.open_rv_12m)', 'MEAN(dat.open_rv_24m)',
      'MEAN(dat.max_bal_bc)', 'MEAN(dat.all_util)',
      'MEAN(dat.total_rev_hi_lim)', 'MEAN(dat.inq_fi)',
      'MEAN(dat.total_cu_tl)', 'MEAN(dat.inq_last_12m)',
      'NUM_UNIQUE(dat.pymnt_plan)', 'MODE(dat.pymnt_plan)'],
      dtype='object', length=332)
```

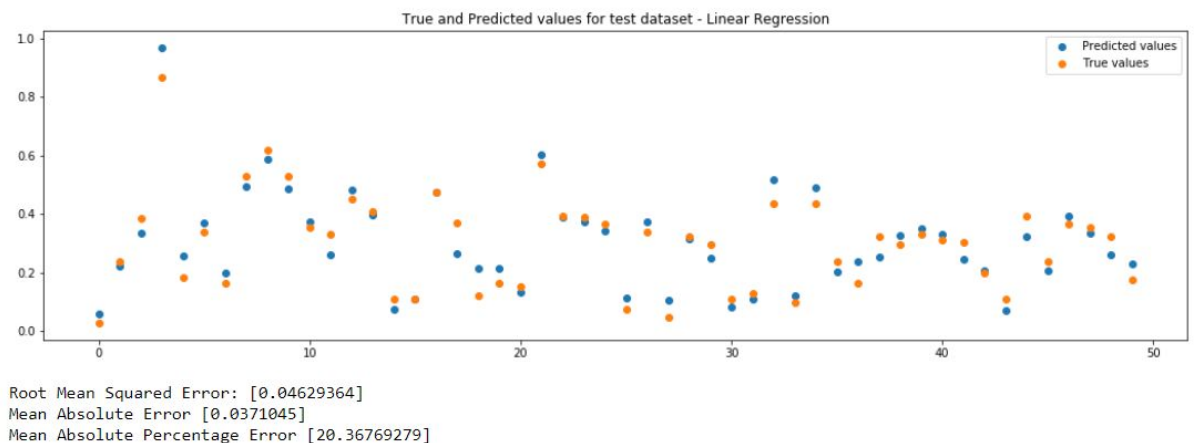
---

## Task 3 - Prediction and MAPE Analysis

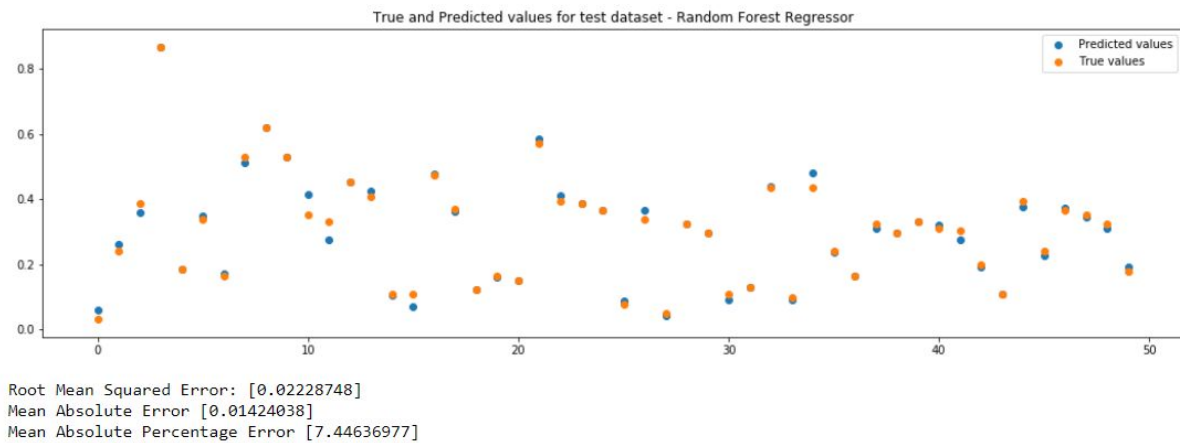
### Train - Test Split

- Used sklearn.model\_selection - train\_test\_split to split data
- Cleaned data split to training and testing data - 80:20

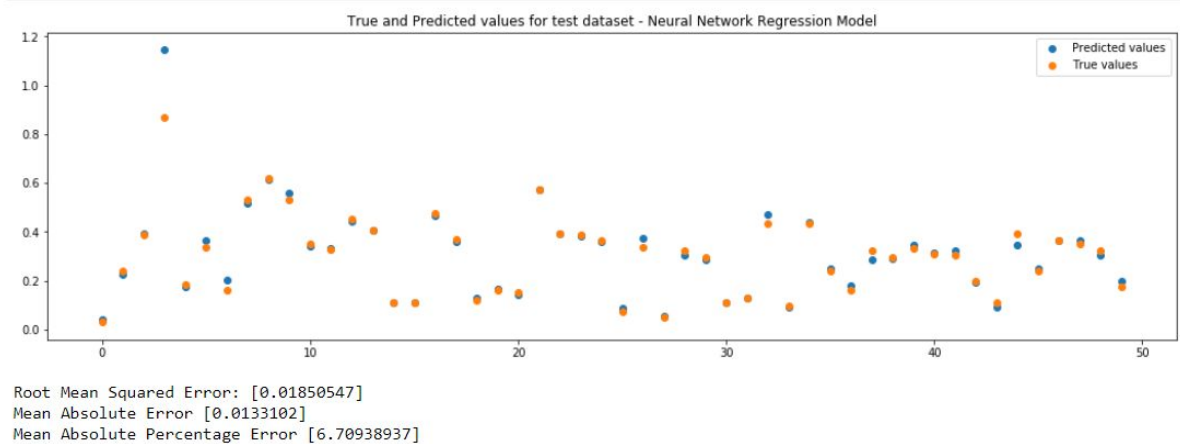
### 1. Linear Regression



## 2. Random Forest



## 3. Artificial Neural Networks



## Model Comparison

MODEL	RMSE	MAE	MAPE
LIN_REG	0.04629	0.03710	20.36769 %
RAND_FRST	0.02229	0.01424	7.44637 %
NN	0.01851	0.01331	6.70939 %

## Model Performance with 5-Fold Cross Validation - Linear Regression

K-Fold CV splits given data into a K number of sections/folds where each fold is used as a testing set at some point.

Here, the data set is split into 5 folds.

In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

Before Cross Validation

```
Root Mean Squared Error: [0.04629364]
Mean Absolute Error [0.0371045]
Mean Absolute Percentage Error [20.36769279]
```

After Cross Validation

```
Root Mean Squared Error: [0.04619486]
Mean Absolute Error [0.03711851]
Mean Absolute Percentage Error [20.36431381]
```

---

## Task 4 - Hyper-parameter Tuning and Auto ML

### Hyper-parameter Optimization

Random Forest

1. Tree depth - max\_depth changed from default (2) to 5
2. Number of trees - changed to 10

Before:

Root Mean Squared Error: [0.02228748]  
Mean Absolute Error [0.01424038]  
Mean Absolute Percentage Error [7.44636977]

After:

Root Mean Squared Error: [0.04161407]  
Mean Absolute Error [0.03435454]  
Mean Absolute Percentage Error [18.13900258]

## Artificial Neural Network

- i) Learning Rate - {'constant', 'invscaling', 'adaptive'}, default 'constant'
- ii) The number of epochs is determined by the maximum number of iterations (max\_iter)

Before:

Root Mean Squared Error: [0.01850547]  
Mean Absolute Error [0.0133102]  
Mean Absolute Percentage Error [6.70938937]

After

Root Mean Squared Error: [0.01778292]  
Mean Absolute Error [0.01289416]  
Mean Absolute Percentage Error [6.90784192]

## Linear Regression

Root Mean Squared Error: [0.04629364]  
Mean Absolute Error [0.0371045]  
Mean Absolute Percentage Error [20.36769279]

Ridge Regression

Root Mean Squared Error: [0.04649018]  
Mean Absolute Error [0.03726418]  
Mean Absolute Percentage Error [20.55064082]

Lasso Regression



```
Root Mean Squared Error: [0.07201117]
Mean Absolute Error [0.05776132]
Mean Absolute Percentage Error [36.96038408]
```

## Elastic Net Regression

```
Root Mean Squared Error: [0.06375659]
Mean Absolute Error [0.05103332]
Mean Absolute Percentage Error [32.89731126]
```

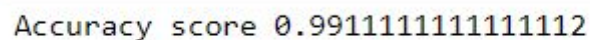
## Auto ML

### H2O.ai

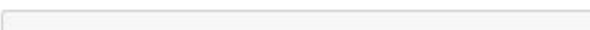
```
MSE: 0.38556212576010623
RMSE: 0.6209364909232716
LogLoss: 1.5478732557099433
Mean Per-Class Error: 0.7679741689552468
Confusion Matrix: Row labels: Actual class; Column labels: Predicted class
```



### Auto SKLearn



```
Accuracy score 0.9911111111111112
```



---

## Task 5 - Testing



sample

Predicted interest rate is: [13.16302049]  
for the following user profile:

3]:

Unnamed: 0		id	loan_amnt	funded_amnt	funded_amnt_inv	term	installment	grade	emp_length
494997	0.55782	0.968446	0.565217	0.565217	0.571429	0.0	0.470391	0.333333	0.9

1 rows × 56 columns

