

Paper BF-236

Know your Interest Rate

Anirban Chakraborty, Dr. Goutam Chakraborty, Oklahoma State University

ABSTRACT

The Federal Reserve of the United States reported a drastic increase in consumer debt over the past few years, reaching \$3.5 trillion in May 2015. Credit card debt accounts for only 26% of total consumer debt, however, the rest of the 74% is derived from student loans, automobile loans, mortgage etc. Lending loans has become an integral part of US consumers' everyday life. Have you ever wondered how lenders use various factors such as FICO score, annual income, the loan amount approved, tenure, debt-to-income ratio, and several others to select your interest rates? The process, defined as 'risk-based pricing', uses a sophisticated algorithm that leverages different determining factors of a loan applicant. This research provides an approach to explore the factors that significantly affect borrowers fixed loan interest rates. For the purpose of this research, data was collected from a publicly available data source – Lending Club, which is the largest peer-to-peer online credit market place. The downloaded dataset has information about successful loan applications in 2015, and includes 421,097 observations and 115 variables. Exploration of data shows that debt consolidation is the primary reason for loan application, comprising 59% of all loan applications followed by credit card bill payments. Further analysis is warranted to determine factors that affect loan interest rate significantly. Selection of significant factors will help develop a prediction algorithm which can estimate loan interest rates based on clients' information. On one hand, knowing the factors will help consumers and borrowers to increase their credit worthiness and place themselves in a better position to negotiate for getting a lower interest rate. On the other hand, this will help lending companies to get an immediate fixed interest rate estimation based on clients' information. By building various predictive models on diverse factors that might influence the interest rate set we take an attempt to answer the following problem statement:

1. Determine if a borrower will receive a low or high interest rate.
2. Estimate the value of the interest rate based on various significant factors.

INTRODUCTION

Lending Club is the world's largest online credit marketplace for peer to peer lending, facilitating personal loans, business loans, and financing for elective medical procedures. Borrowers access these loans through fast and easy online or mobile interfaces. Investors provide the capital to enable many of the loans in exchange for earning interest. Peer-to-peer (P2P) lending is the practice of lending money to individuals or businesses through online services that match lenders directly with borrowers. Since the P2P lending companies offering these services operate entirely online, they can run with lower overhead and provide the service more cheaply than traditional financial institutions. As a result, lenders often earn higher returns compared to savings and investment products offered by banks, while borrowers can borrow money at lower interest rates, even after the P2P lending company has taken a fee for providing the match-making platform and credit checking the borrower. The interest rates are set by lenders who compete for the lowest rate on the reverse auction model or fixed by the intermediary company on the basis of an analysis of the borrower's credit.

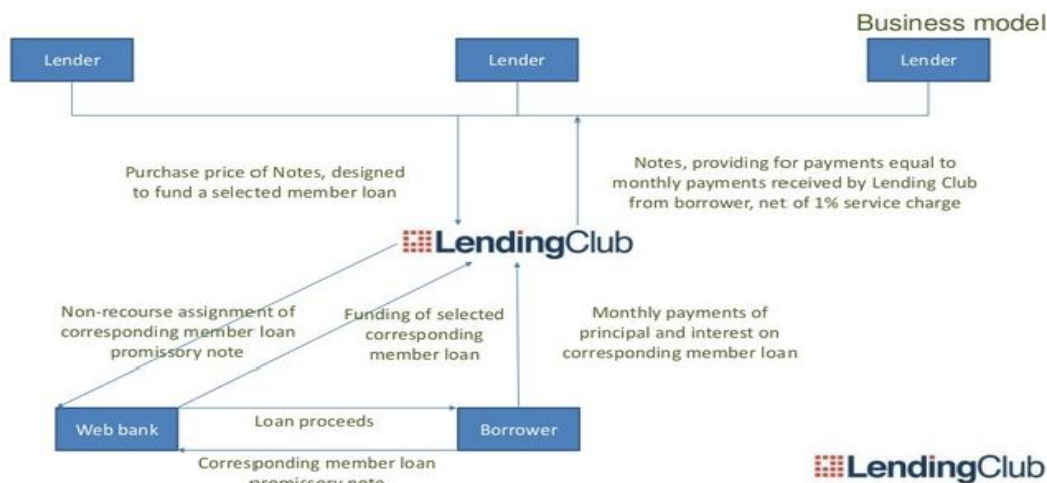


Fig. 1. Lending Club's Business Model

The process of an application for a loan to its approval can be explained primarily in three steps:

- Customers interested in a loan complete a simple application at LendingClub.com.

- Lending Club leverages online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates. Qualified applicants receive offers and can evaluate loan options with no impact to their credit score.
- Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns.

DATA DESCRIPTION & METADATA

The loan data for 2015 was collected from the Lending club website and has 421097 observations and 115 variables. Embedded is the data dictionary:



LC Loan data dictionary

METHODOLOGY

The downloaded datasets were in the form of CSV files, which were converted into SAS Datasets. After exploratory analysis, various predictive models are built keeping various factors that influence the interest rate as independent variables and the interest rate as the dependent variable. The aim of the project is to Estimate the interest rate.

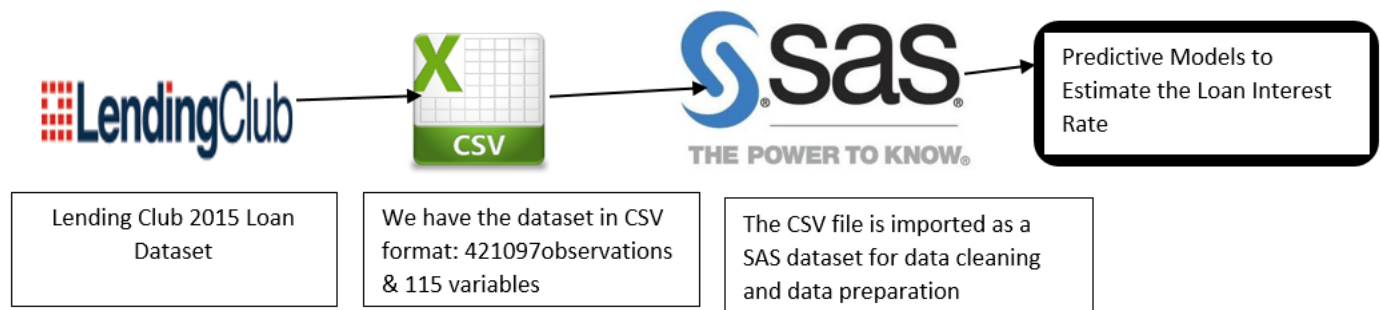


Fig. 2. Project Methodology

We took a two-way approach in determining the interest rates, first we determined if the interest rate was going to be low or high and then the value of interest rates within each category.

DATA PREPARATION & EXPLANATORY ANALYSIS

The data was then taken in to SAS Enterprise Miner and was put under cleaning phase that did things such as: 1) removing the variables with large missing values, 2) correcting the skewness in the data by data transformation. A few variables had missing values which were imputed using tree imputation.



Fig 3. Shows the distribution of all the 2015 loan borrowers across all the states. We can see that the majority of borrowers are from the states: California, Texas, New York and Florida respectively. Of all the borrowers there was no one from the states Idaho and Iowa maybe because of the state laws here.



Fig 4. Indicates the most common reasons for loan applications were: Debt Consolidation (59%), Credit Card, home improvement or a major Purchase.

We took a two-way approach in determining the interest rate:

To Predict the Interest Rate Category

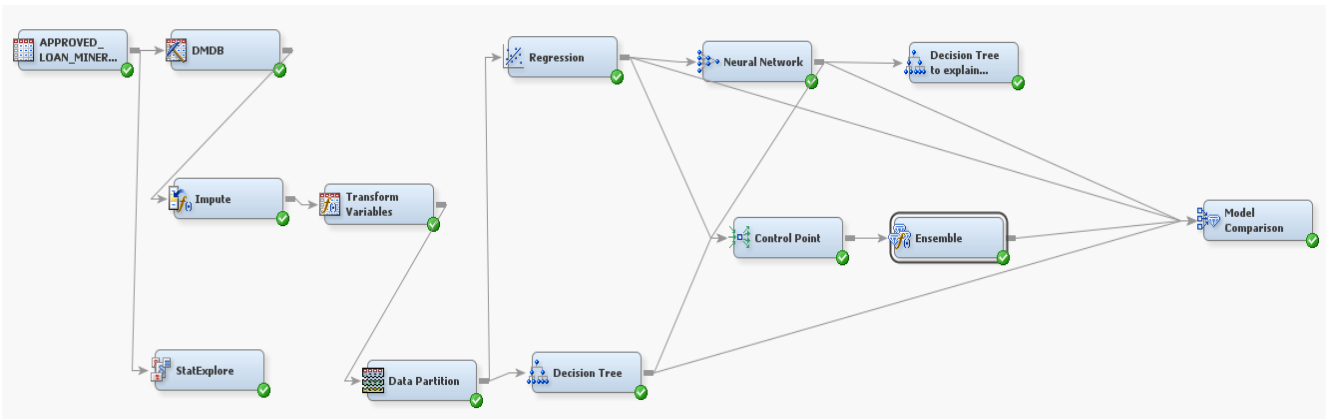


Fig. 5. Modelling Layout for Predicting the Loan Interest Rate Category

High level steps for the above modelling are as follows:

- **STEP 1:** From the data we found out that the median interest rate was around 15%, any predicted interest rate below 15% was categorized as low interest rate (predicted output =1) and predicted interest rates above 15% as high interest rate (predicted output =0).
- **STEP 2:** After data cleaning was done, we built several models like logistic regression, neural network, decision tree, an ensemble model of the previous three and gradient boosting model, to best predict the category of the interest rate for all the borrowers.
- **STEP 3:** The neural network model was conducted via stepwise logistic regression for variable selection.
- **STEP 4:** From Model Comparison we found that, neural network best predicts the category of the interest, which is explained by back tracking the results into a decision tree.

Process Flow

The raw dataset consisted of 116 variables and 421097 observations which was further reduced to 80 variables as the rejected 26 variables were not applicable to (predicting interest rate) and most of the removed variables had around 70% - 80% missing values.

Variable Summary

Role	Measurement Level	Frequency Count
ID	NOMINAL	1
INPUT	INTERVAL	69
INPUT	NOMINAL	6
REJECTED	INTERVAL	1
REJECTED	NOMINAL	3
TARGET	BINARY	1

Fig. 6. Variable Summary Before Variable Imputation

The reduced data set was then taken into SAS Enterprise Miner platform, using 'StatExplore' node, the distribution of the variables was checked in order to clean the data by imputing the missing values and normalizing any variables with high skewness. Finally, the data set was converted to a clean data source. As a part of CRISP-DM methodology data was then partitioned as training (70%) and validation (30%). Once the data was split, various modelling techniques were used in order to predict the category (High/Low) of the interest rate. We had to reject variables like INTEREST_CLASS, INTEREST_SUBCLASS, MONTHLY_INTEREST_AMOUNT and LOAN_ID as these were a part of the desired output.

Regression Model: Logistic regression was applied on the training data, with stepwise selection as the variable selection model and validation misclassification rate as the model selection criterion (decision output).

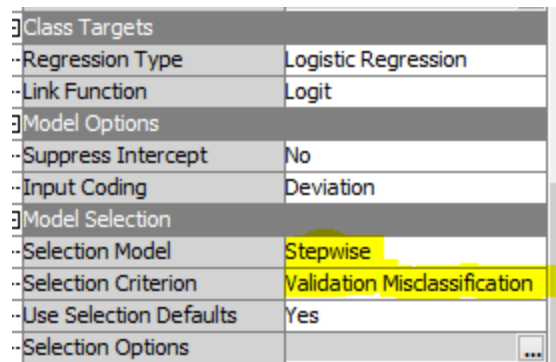


Fig. 7. Logistic Regression Properties

We also kept the imputed indicator variables (10 imputed indicator variables), so that the model correctly knows if the observation has a true value or an imputed value which might in turn have an effect in the prediction.

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	10
INPUT	INTERVAL	69
INPUT	NOMINAL	6
REJECTED	INTERVAL	1
REJECTED	NOMINAL	3
TARGET	BINARY	1

Fig. 8. Variable Summary After Data Preparation

The Regression model went through 48 steps before obtaining the final solution. The prediction of the high/low interest rate was based on the following forty-two variables with high level of significance:

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
IMP_inq_last_6mths	1	1720.1289	<.0001
IMP_mo_sin_old_il_acct	1	420.1111	<.0001
IMP_mths_since_recent_bc	1	228.6511	<.0001
IMP_mths_since_recent_inq	1	316.8081	<.0001
IMP_percent_bc_gt_75	1	1093.5081	<.0001
LOG_acc_now_delinq	1	96.2819	<.0001
LOG_annual_inc	1	1894.4642	<.0001
LOG_collection_recovery_fee	1	5.7078	0.0169
LOG_dti	1	558.7760	<.0001
LOG_revol_bal	1	41.7838	<.0001
M_bc_util	1	474.9581	<.0001
M_mo_sin_old_il_acct	1	160.7071	<.0001
acc_open_past_24mths	1	341.6500	<.0001
addr_state	49	140.3204	<.0001
application_type	2	1225.6489	<.0001
credit_history	1	38.2706	<.0001
delinq_2yrs	1	95.8319	<.0001
emp_length	11	261.1684	<.0001
fico_range_low	1	1256.9390	<.0001
home_ownership	2	66682.7402	<.0001
last_fico_range_high	1	1167.2784	<.0001
last_pymnt_amnt	1	13239.3394	<.0001
loan_amnt	1	160.1530	<.0001
mo_sin_old_rev_tl_op	1	278.4017	<.0001
mo_sin_rcnt_rev_tl_op	1	34.7725	<.0001
num_bc_sats	1	23.1348	<.0001
num_bc_tl	1	130.0786	<.0001
num_tl_op_past_12m	1	2740.5241	<.0001
open_acc	1	60.7436	<.0001
out_prncp	1	506.5850	<.0001
out_prncp_inv	1	516.0265	<.0001
purpose	13	8494.0340	<.0001
term	1	202.9233	<.0001
total_acc	1	315.6666	<.0001
total_bal_ex_mort	1	431.2780	<.0001
total_bc_limit	1	409.7188	<.0001
total_il_high_credit_limit	1	325.3076	<.0001
total_pymnt	1	356.0643	<.0001
total_pymnt_inv	1	385.3574	<.0001
total_rec_int	1	268.0130	<.0001
total_rec_prncp	1	23.0008	<.0001
total_rev_hi_lim	1	199.6864	<.0001

Fig. 9. Significant Variables Selected After Stepwise Logistic Regression

Fit Statistics	Statistics Label	Train	Validation
AIC_	Akaike's Information Criterion	246086.4	.
ASE_	Average S Statistics Label	0.091181	0.091553
AVERR_	Average Error Function	0.316392	0.317974
DFE_	Degrees of Freedom for Error	388358	.
DFM_	Model Degrees of Freedom	129	.
DFT_	Total Degrees of Freedom	388487	.
DIV_	Divisor for ASE	776974	332990
ERR_	Error Function	245828.4	105882.3
FPE_	Final Prediction Error	0.091242	.
MAX_	Maximum Absolute Error	1	1
MSE_	Mean Square Error	0.091212	0.091553
NOBS_	Sum of Frequencies	388487	166495
NW_	Number of Estimate Weights	129	.
RASE_	Root Average Sum of Squares	0.301962	0.302577
RFPE_	Root Final Prediction Error	0.302063	.
RMSE_	Root Mean Squared Error	0.302012	0.302577
SBC_	Schwarz's Bayesian Criterion	247488.6	.
SSE_	Sum of Squared Errors	70845.43	30486.15
SUMW_	Sum of Case Weights Times Freq	776974	332990
MISC_	Misclassification Rate	0.128071	0.128058

Fig. 10. Logistic Regression Fit Statistics

Fig 10. shows that the misclassification rate is 12.8% for the validation data.

Decision Tree: Decision trees have a variety of outputs that we can use. From estimating variable importance to scoring a new data set, they can be used for a variety of purposes. An important point to note here was the assessment measure chosen for the stopping point of the tree is misclassification rate. We chose to use this misclassification rate because the end goal of this model was to provide a decision of high or low interest rate. This

decision could be used in knowing the different criteria required to receive a low or a high loan interest rate. Also, please note the other options like max branch, max depth etc. are good to work with for optimizing complexity of the decision tree.

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25

Fig. 11. Decision Tree Properties

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
total_rec_int		15	1.0000	1.0000	1.0000
term		1	0.9628	0.9782	1.0159
total_rec_prncp		16	0.8538	0.8438	0.9884
fico_range_high		1	0.5231	0.5232	1.0002
last_fico_range_high		1	0.3954	0.3886	0.9828
LOG_dti	Transformed dti	2	0.2867	0.2708	0.9445
total_pymnt_inv		2	0.1959	0.1902	0.9711
application_type		1	0.1323	0.1228	0.9281
last_pymnt_amnt		1	0.0985	0.0966	0.9813

Fig. 12. Decision Tree Variable Importance

We found that the most important variables in predicting the interest category were interest amount received till date, the loan tenure followed by total principal received till date and the current FICO score (higher score) and other variables such as previous FICO score (lower score), debt to income ratio, application type for loan and the last installment amount .

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	388487	166495
MISC	Misclassification Rate	0.145168	0.14574
MAX	Maximum Absolute Error	0.996109	0.996109
SSE	Sum of Squared Errors	89071.68	38303.64
ASE	Average Squared Error	0.114639	0.115029
RASE	Root Average Squared Error	0.338584	0.33916
DIV	Divisor for ASE	776974	332990
DFT	Total Degrees of Freedom	388487	.

Fig. 13. Decision Tree Fit Statistics

Fig 13. Shows the misclassification rate achieved for by the decision tree model is around 15%.

Neural Network: it is a network inspired by biological neural networks (the brain) which are used to estimate or approximate functions that can depend on a large number of inputs that are generally unknown. A typical network

has hidden layers that consist of neurons to which the input variables are received and the network processing results is the desired output. These neural models are also referred as black box models and hence are very complicated to explain, however with the help of a decision tree as a backpropagation the neural network model can be explained. The variable selection was done via stepwise logistic regression.

Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	3
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Fig. 14. Neural Network Properties

The architecture of the neural network was taken as Multilayer Perceptron(feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs) and 3 hidden units were chosen(default settings of neuralnetwork model).

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
term		1	1.0000	1.0000
total_rec_int		4	0.6255	0.6189
total_rec_prncp		4	0.5652	0.5431
fico_range_low		1	0.5433	0.5349
last_fico_range_high		1	0.4107	0.3973

Fig. 15. Neural Network Variable Importance

From the neural network model , we see that the most important variables in determining whether a person will get a low or high interest are : term of loan, interest amount received till date, principal loan amount received till date, current lower FICO score followed by previous higher FICO score.

Data Role=VALIDATE Target=category Target Label=' '

False	True	False	True		
Negative	Negative	Positive	Positive	Sensitivity	Specificity
				$\frac{TP}{TP + FN} = 92.48\%$	$\frac{TN}{TN + FP} = 47.43\%$
9356	19915	22069	115155		

Fig. 16. Sensitivity & Specificity

Fig 16. Indicates that the neural net model measures the proportion of positives that are correctly identified (92.48% times) and the proportion of negatives that are correctly identified (47.43%).

```

-----*
Node = 14
-----*
if total_rec_int < 2485.1 or MISSING
AND term IS ONE OF: 36 MONTHS or MISSING
AND last_fico_range_high >= 676.5 or MISSING
then
    Tree Node Identifier    = 14
    Number of Observations = 186748
    Predicted: category=1 = 0.90
    Predicted: category=0 = 0.10

```

Fig. 17. Sample English Rule

Fig 17. Shows one of the English rules for the prediction of low or high Interest rate. As per node 14 if a borrower has already paid interest amount up to \$2,485, has a loan tenure of 36 months and has a FICO score of 676.5 or higher at last check, chances are the borrower will get a lower interest rate of less than 15%.

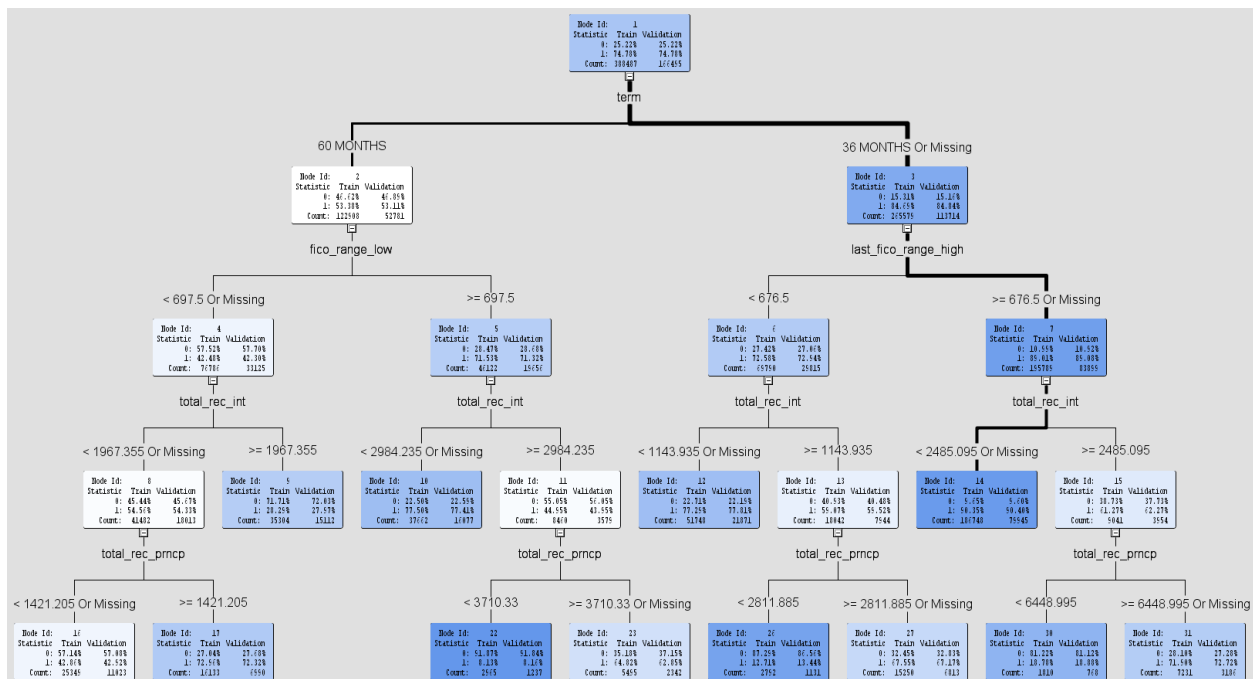


Fig. 18. Neural Network as Explained by The Back Tracking Tree Structure

Following were the findings from the Tree Structure:

- We set the number of branches as two and the number of levels as four, as a part of the pruning process. From the above tree we can see that the model gives four major splits. The first split occurs for the variable loan tenure, if it is for 36 months or 60 months.
- The second split occurs for previous lower fico score and current upper fico score.
- Next the model splits on the basis of total interest amount received till date followed by the total principal loan amount received till date.
- The nodes in blue depicts that the decisions as per these nodes have high significance and the results show higher confidence.

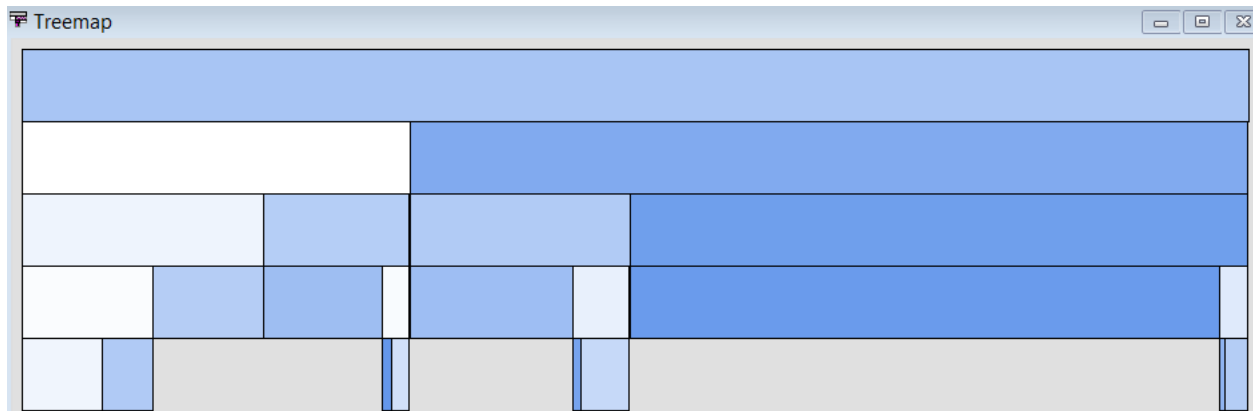


Fig. 19. Tree Map

Looking at the decision tree map shown above, we get a very compact overview of the separation between the low interest rate and high interest rate. The node width is representative of the number of observations contained in the node after the split. Also, by default, the color of the node is reflective of the distinct difference of the split between high/low interest rate i.e. the darker the color the better is the split between the high/low interest rate.

On comparing all the models, we found that, neural network best predicts the category of the interest rate compared to regression or decision tree as it resulted in lower misclassification rate of 7%.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Neural	Neural	Neural Net...	category		0.070086
	Ensmbl	Ensmbl	Ensemble	category		0.092627
	Reg	Reg	Regression	category		0.128058
	Tree	Tree	Decision Tr...	category		0.14574

Fig. 20. Model Comparison

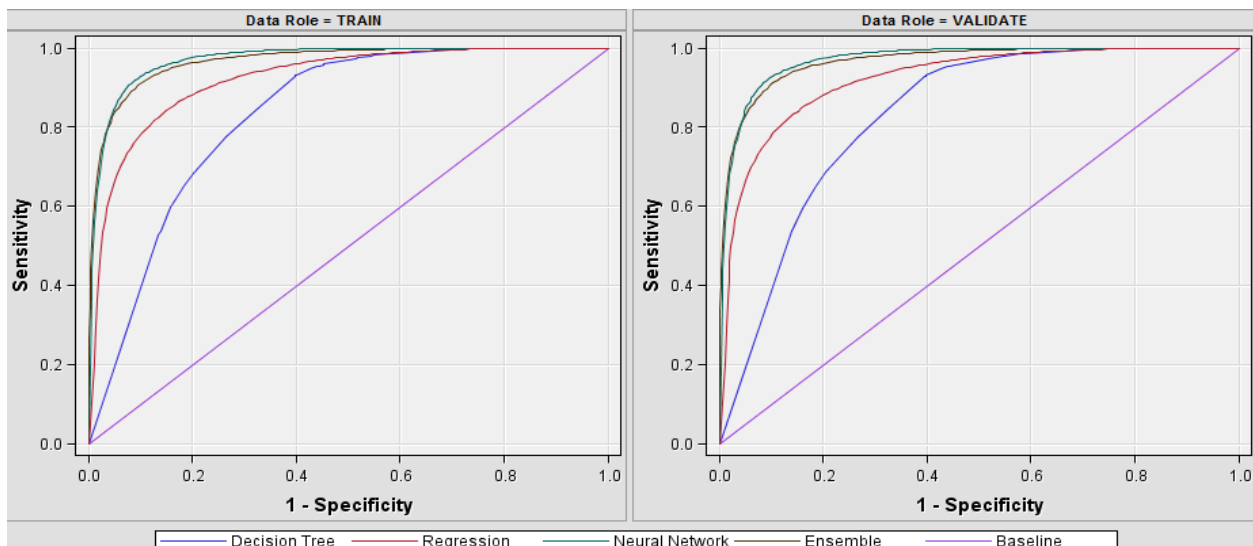


Fig. 21. Model Comparison ROC Curves

ROC curve, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. Neural network has the best ability of the test to discriminate between the two possible outcomes high/low interest rate as shown in Fig 21.

To Predict the Interest Rate Value

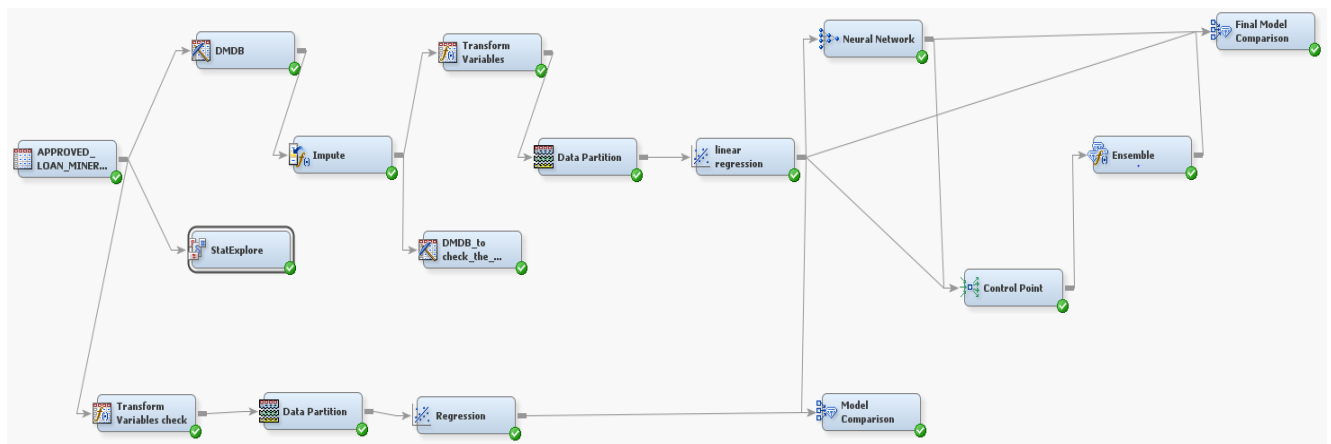


Fig. 22. Modelling Layout for Predicting the Value of Loan Interest

High level steps for the above modelling are as follows:

- **STEP 1:** After the data was prepared, we built three models i.e. Liner regression, Neural Network and an ensemble of the previous two to predict the interest rate.
- **STEP 2:** The variable selection neural network model was done via stepwise linear regression.
- **STEP 3:** The Neural Network model best predicts the interest rate, as it has the lowest average squared error. From the fit statistics of the regression model we can speak about the performance of the model, it has an adjusted R squared value of 0.6217.

Process Flow

The raw data source like previous modelling scenario went through the required imputation and transformation. Additionally, a similar kind of split between the training and validation was done before building different predictive models on the training data. A point to be noted here is that the target was now set as interval variable (INTEREST_RATE) as were predicting the value of the interest rate of any borrower.

Variable Summary		
Role	Measurement Level	Frequency Count
ASSESS	NOMINAL	1
ID	NOMINAL	1
INPUT	BINARY	10
INPUT	INTERVAL	69
INPUT	NOMINAL	6
REJECTED	NOMINAL	3
TARGET	INTERVAL	1

Fig. 23. Variable Selection After Data Preparation

Linear Regression: The first modelling technique applied was liner regression, as it is a pretty robust modelling technique.

Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

Fig. 24. Linear Regression Properties

Stepwise selection was used as a variable selection method and the validation error was set as the assessment method because here we are trying to estimate the range of the interest rate.

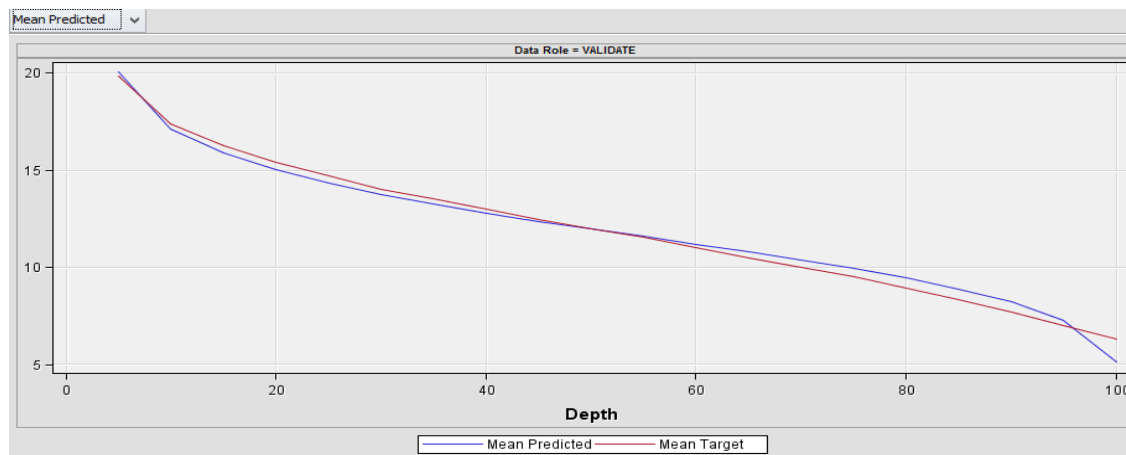


Fig. 25. Mean Predicted Vs Mean Target

We can see in Fig 25. That the plot of mean predicted and mean target follow similar pattern, which tells how closely the mean predicted values are to the mean target values.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	107	4691076	43842	5967.02	<.0001
Error	388379	2853560	7.347358		
Corrected Total	388486	7544636			

Model Fit Statistics

R-Square 0.6218 Adj R-Sq 0.6217

Fig. 26. Linear Regression Adjusted R Square

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
interest_rate		ASE	Average Squared Error	7.274635	7.280425

Fig. 27. Linear Regression Average Square Error

R squared is a statistic that gives information about the goodness of fit of a model. In regression, the adjusted R squared coefficient is a statistical measure of how well the regression line approximates the real data points. In our model we have around 62% adjusted R squared value, with an average error of around 7.3

Neural Network: The regression node was connected to a neural network model as a part of variable selection (stepwise method).

Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No

Fig. 28. Neural Network Properties

Following were some of the significant variables, which helped in predicting the loan interest rate.

Type 3 Analysis of Effects				
Effect	DF	Sum of Squares	F Value	Pr > F
IMP_inq_last_6mths	1	42370.4025	5766.75	<.0001
IMP_mo_sin_old_il_acct	1	13871.5959	1887.97	<.0001
IMP_mths_since_recent_bc	1	3995.1463	543.75	<.0001
IMP_mths_since_recent_inq	1	5713.6113	777.64	<.0001
IMP_percent_bc_gt_75	1	32890.4134	4476.49	<.0001
IMP_revol_util	1	2656.0220	361.49	<.0001
LOG_acc_now_delinq	1	3417.1753	465.09	<.0001
LOG_annual_inc	1	36719.1082	4997.59	<.0001
LOG_collection_recovery_fee	1	4571.7604	622.23	<.0001
LOG_dti	1	3172.6569	431.81	<.0001
M_mo_sin_old_il_acct	1	2567.5281	349.45	<.0001
M_percent_bc_gt_75	1	9401.0301	1279.51	<.0001
acc_open_past_24mths	1	14126.4664	1922.66	<.0001
addr_state	49	2897.9563	8.05	<.0001
application_type	2	58454.6428	3977.94	<.0001
avg_cur_bal	1	2788.0073	379.46	<.0001
delinq_2yrs	1	2714.4698	369.45	<.0001
emp_length	11	9867.4510	122.09	<.0001
fico_range_low	1	49110.7904	6684.14	<.0001
home_ownership	3	8335.5883	378.17	<.0001
last_fico_range_high	1	48143.4145	6552.48	<.0001
last_pyamt_amt	1	205037.156	27906.2	<.0001
mo_sin_old_rev_tl_op	1	15897.2127	2163.66	<.0001
num_bc_tl	1	6932.7668	943.57	<.0001
num_tl_op_past_12m	1	59646.3862	8118.07	<.0001
purpose	13	226250.534	2368.73	<.0001
term	1	235835.179	32098.0	<.0001
total_acc	1	15867.9859	2159.69	<.0001
total_bal_ex_mort	1	10903.4239	1483.99	<.0001
total_bc_limit	1	8475.7027	1153.57	<.0001
total_il_high_credit_limit	1	8372.1520	1139.48	<.0001
total_rec_int	1	677135.789	92160.4	<.0001
total_rec_prncp	1	167511.169	22798.8	<.0001
total_rev_hi_lim	1	3515.7990	478.51	<.0001

Fig. 29. Neural Network Significant Variables

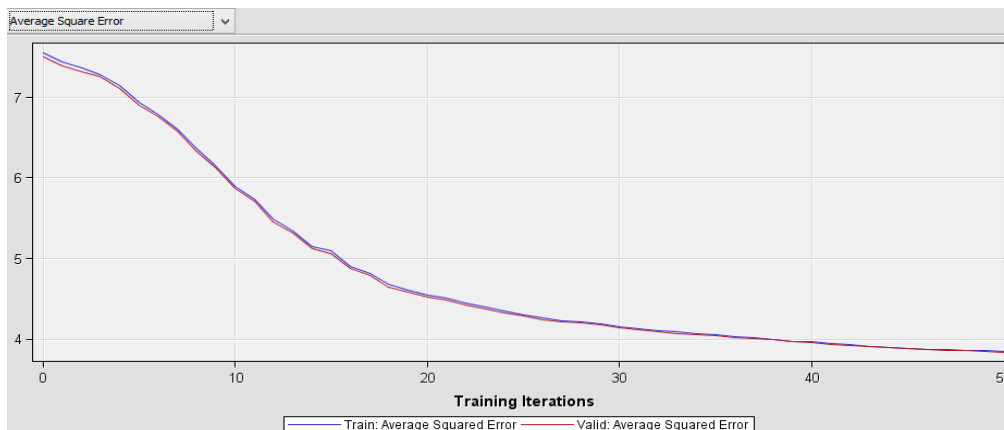


Fig. 30. ASE Training Vs Validation

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
interest_rate		_ASE_	Average Squared Error	3.845458	3.842557

Fig. 31. Neural Network average square error

We can see that the, training and validation data are closely bound and have similar kind of error function, which reduces significantly as the number of iterations increases over time. The validation data has an average squared error of 3.84.

Data Role=TRAIN Target Variable=interest_rate Target Label=' '					Data Role=VALIDATE Target Variable=interest_rate Target Label=' '				
Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score	Range for Predicted	Mean Target	Mean Predicted	Number of Observations	Model Score
24.592 - 25.804	24.3656	25.1900	1920	25.1980	24.552 - 25.804	24.3436	25.1551	873	25.1783
23.379 - 24.592	23.1432	23.9563	2255	23.9852	23.301 - 24.552	23.0252	23.8980	1030	23.9266
22.166 - 23.379	22.1708	22.7410	2875	22.7723	22.049 - 23.301	22.2017	22.6357	1314	22.6749
20.953 - 22.166	21.0956	21.5272	3963	21.5595	20.797 - 22.049	20.9359	21.3823	1842	21.4232
19.740 - 20.953	19.9470	20.3057	5774	20.3467	19.546 - 20.797	19.8333	20.1201	2676	20.1715
18.527 - 19.740	18.9766	19.0901	8801	19.1339	18.294 - 19.546	18.8103	18.8702	4097	18.9198
17.315 - 18.527	18.0228	17.8847	12828	17.9210	17.042 - 18.294	17.8131	17.6345	6131	17.6681
16.102 - 17.315	17.0826	16.6826	16324	16.7082	15.791 - 17.042	16.8185	16.3907	7669	16.4164
14.889 - 16.102	15.8458	15.4619	21188	15.4954	14.539 - 15.791	15.4653	15.1257	10096	15.1647
13.676 - 14.889	14.5102	14.2362	30698	14.2826	13.287 - 14.539	14.1886	13.8517	16042	13.9130
12.463 - 13.676	13.3786	13.0533	53218	13.0697	12.035 - 13.287	12.9171	12.7103	24435	12.6613
11.251 - 12.463	11.8485	11.8638	47276	11.8569	10.784 - 12.035	11.4122	11.4145	20489	11.4096
10.038 - 11.251	10.6073	10.6477	45220	10.6441	9.532 - 10.784	10.0655	10.1611	19622	10.1579
8.825 - 10.038	9.2761	9.4322	43992	9.4313	8.280 - 9.532	8.7444	8.9035	19368	8.9062
7.612 - 8.825	7.9458	8.1907	47851	8.2184	7.029 - 8.280	7.3780	7.7295	18446	7.6545
6.399 - 7.612	6.8017	7.0816	25965	7.0056	5.777 - 7.029	6.4458	6.4778	7963	6.4028
5.186 - 6.399	6.1005	5.8648	13009	5.7928	4.525 - 5.777	5.8178	5.2567	3656	5.1511
3.974 - 5.186	5.7568	4.7564	4704	4.5800	3.274 - 4.525	6.4404	4.1261	663	3.8994
2.761 - 3.974	7.0884	3.5102	577	3.3671	2.022 - 3.274	6.9259	2.9512	81	2.6477
1.548 - 2.761	7.7755	2.5352	49	2.1543	0.770 - 2.022	11.0000	1.2566	2	1.3960

Fig. 32. Assessment Score Distribution

Fig 32. Shows the scoring distribution of the range of Interest rate by the neural network model for the training and the validation data.

Ensemble Model: is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy. We decided to create an ensemble model having both the regression and the neural network model.

Fit Statistics				
Target=interest_rate Target Label=' '				
Fit Statistics	Statistics Label	Train	Validation	
ASE	Average Squared Error	4.64	4.63	

Fig. 33. Ensemble Model Fit Statistics

On comparing all the three different models, we find that the neural network model performs the best in predicting the interest rate.

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Train: Average Squared Error
Y	Neural	Neural	Neural Net...	interest_rate		3.845458
	Ensmbl	Ensmbl	Ensemble	interest_rate		4.640659
	Reg	Reg	linear regre...	interest_rate		7.355667

Fig. 34. Model Comparison

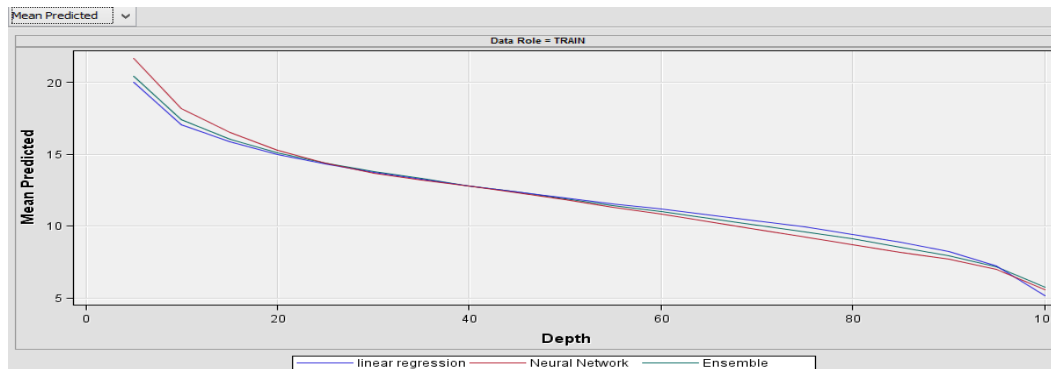


Fig. 35. Mean Predicted Comparison

CONCLUSION

It is advisable for any borrower to maintain a good lower fico credit score on the last update credit history and have a high upper credit score in the current credit check in order to receive lower interest rates. The amount of loan amount doesn't play a significant role while deciding the interest rate, chances are a person with high loan amount might receive a lower interest rate compared to a person who has minimal loan amount. Borrowers with employment more than 7 years tend to pay lower interest rates, compared to other people. The states from where a loan is being applied also plays a pivotal role in determining the interest rates, states like Washington, Virginia, Tennessee, Ohio, Minnesota, New York and Illinois fall in this category. If a borrower owns a house, chances are he will pay lower interest rates compared to people who stay in a rented house. The Income and debt-to-income ratio of a borrower plays a significant role in determining the interest rate.

FUTURE WORK

The scope of this model can be extended by bringing in macro-economic variables like the inflation rate and the average annual income across all states, in determining the interest rates. Also it would be worth studying the effect of the stock exchange index like S&P 500, NASDAQ Composite, DOW Jones Industrial Average etc. effect on fluctuation in the interest rates.

REFERENCES

- [1] Massimo Guidolin, Allan Timmermann , Forecasts of US Short-term Interest Rates
- [2] James E. Pesando, Forecasting Interest rates: an efficient Markets perspective
- [3] Duffee, Gregory H. 2013. "Forecasting Interest Rates.", Handbook of Economic Forecasting Vol. 2
- [4] Arito Ono, Kosuke Aoki, Shinichi Nishioka, Kohei Shintani, Yosuke Yasui, Long-term interest rates and bank loan supply
- [5] <https://www.lendingclub.com/info/download-data.action>
- [6] <https://www.wikipedia.org/> --- for datamining definitions

ACKNOWLEDGEMENT

- I heartily thank Dr. Goutam Chakraborty, Professor, Department of Marketing and founder of SAS and OSU Data Mining Certificate program, Director MS in Business Analytics – Oklahoma State University for his constant inputs and guidance without which I couldn't have done this research project.
- I would also like to express my gratitude to Dr. Miriam McGaugh, Professor, Department of Marketing – Oklahoma State University for reviewing the research work done and for providing valuable suggestions.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Anirban Chakraborty

Oklahoma State University

Email: anirban.chakraborty@okstate.edu

Anirban Chakraborty is a graduate student currently pursuing Masters in Business Analytics at the Spears School of Business, Oklahoma State University. He is currently working research assistant at Oklahoma State University with Love's Travel Stop under the mentorship of Dr. Goutam Chakraborty. He is SAS® Certified programmer and SAS® Certified Statistical Business Analyst.

Dr. Goutam Chakraborty
Oklahoma State University

Email: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS and OSU data mining certificate and SAS and OSU marketing analytics certificate at Oklahoma State University. He has published many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has over 25 Years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.