Strategizing P2P Investments using Socio-Economic Factors

Priyabrata Thatoi, Oklahoma State University

ABSTRACT

A traditional banking counterpart, Lending Club is an online market place for borrowers and lenders and has become a major player in the peer-to-peer lending business with an average rate of return between 3% and 8% per year. Lenders, before committing their money, carefully investigate a multitude of associated risks such as borrower defaults, liquidity risk, poor loan diversification, etc. But for a steady return, lenders are advised to diversify their portfolio by investing in different loans with different rate of interest. Also, several studies have been conducted with the help of sophisticated machine learning algorithms and traditional credit risk modeling or credit scoring to understand the potential indicators to loan default, which is a major cause of poor return. Most of these modeling techniques utilize collected borrower's personal, professional and credit information. This is crucial for non-traditional banks such as Lending Club because to maintain a low-interest rate and expand their customer base, they need to accurately identify and decline potential defaulters. One factor that is rarely assessed over the period is the borrower's zip-code specific socio-economic indicators such as no. of workers, workers in different industry & occupation, below poverty line families, unemployment rate, etc. in a location. In the latest study, median household income for individual zip-code was used to predict loan default and has reported an increase in the accuracy [3]. This study attempts to strategize investment portfolio by two-stage scoring approach, which is an integration of classification of default loans and prediction of annualized rate of return (ARR) using zip-code specific socio-economic indicators from American Fact Finder and loan history data from Lending Club. More specifically, this study will explore re-sampling techniques such as random under-sampling, random over-sampling and SMOTE to further increase the accuracy in prediction using SAS® Enterprise Miner™, 7 SAS® Enterprise Guide®, and Python 3. The numerical study indicates the predicted return and portfolio size is more realistic and better than existing investment methods such as bonds, savings accounts & Prosper Lending.

INTRODUCTION

Lending Club is an easy to access peer to peer (P2P) lending platform for lenders and borrowers that have become popular due to lower overhead costs, the lower rate of interest, lower penalties, etc. In 2012, the company became the largest lender in the United States based on issued loan volume and revenue. In 2019, they issued about $38 billion with a solid annual return of 3-8%. The process of lending starts with a borrower filing a loan application that goes through a series of screening and an interest rate that is decided using an in-house proprietary model. Approved loans are then listed on the platform for a registered investor to fund by purchasing Lending club notes. These notes are nothing but the assets that correspond to a fraction of loans and a lender can either fund entirely or partially depending upon his lending acumen. Predicting the rate of return has been a tough task as the dynamics behind its change over time. But in general, Investors evaluate information available to them in their investment account such as loan amount, loan duration, borrower credit score, income to debt ratio, risk tolerance, and time horizon, etc. In theory, lenders seem to bet more on safe loans with less return rather than risky loans with high return. Lending Club's derived attribute: grade is an indicator of safe and risky loans. Every loan that is listed varies in grade and interest rates and higher the interest rate, risker the grade and higher the chances of the loan being defaulted. Once the borrower fails to repay the debt, the lender loses the money. There are many potential solutions to avert this, but Lending Club has stressed more on diversification of investment, that is funding partially and varieties of loans.

But why not explore historical loan data available at the lending club and understand the true nature of defaulted borrowers in predicting the rate of return per year? Ability to repay the debt is generally affected by the depreciation in the financial condition of a borrower. If socio-economic factors such as GDP, unemployment rate, GNP, etc. determines the financial health of the country, then it is also logical to link financial condition of a borrower to the socio-economic factors of the location to which the borrower belongs to. Lending club's historical loan data have three-digit zip code and prior projects like "Predicting default risk of Lending Club Loans" have used median household income and population [2] for individual zip code in predicting loan default rate. Also, in "The sensitivity of the loss given default rate to systematic risk" [2, 3], the author has given proof of linkage between default rate and macroeconomic factors [2]. This project aims to classify if a loan is going to be charged off or fully paid and predict Annualized rate of return using socio-economic features such as mean income, population, ratio of private to government workers, total workers, and percentage of people in below poverty line family etc. along with Lending club loan features available to an investor at the time of investment. Socio-economic indicators for individual zip code is available at American Fact Finder. Finally, in an attempt to reduce the volatility in return, the average expected the annualized rate of return was calculated for increasing portfolio size.

KEY FACTORS AND DATA EXPLORATION

Lending Club loans have 60 and 36 months term and four stages of loan status that change from current to either Fully Paid or Charged Off. For a loan to change its status from Current to Late, the payment must due between 16 to 120 days. If the payment dues more than 121 days [4], the loan status changes to Default. Once the loan is the default, then Lending club changes the status to Charged off in the next 2 to 3 weeks. This concludes that a loan takes about 5 months in addition to the loan end date to completely expire.

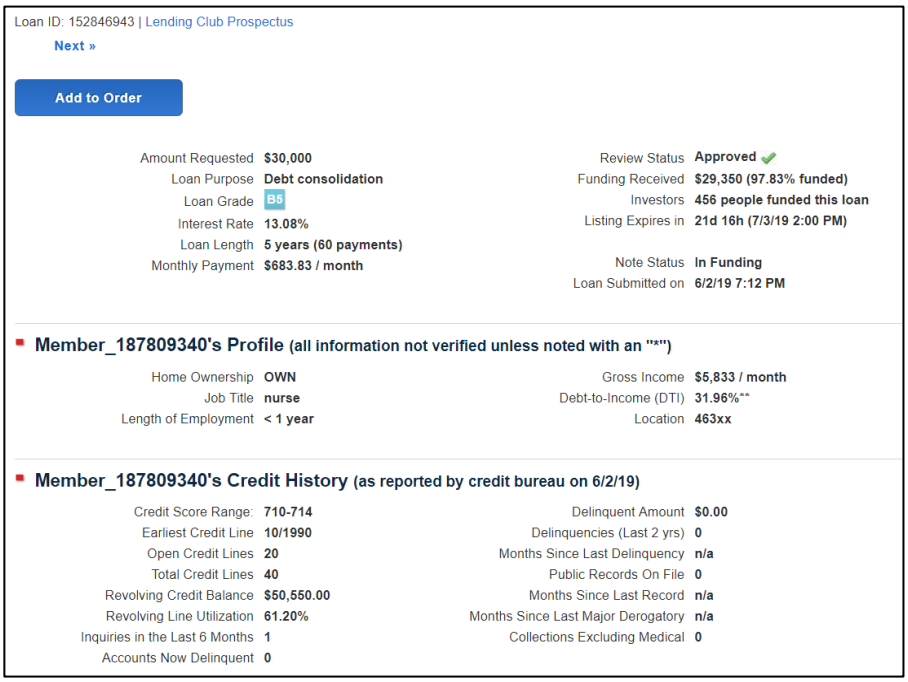

Figure 1. List of variables in a loan listing available to the investor

This study will only consider the loans that have completely expired. Lending club data for the year 2015 has 150 columns and 421095 expired loan ids. However, only a few variables are available to the investor as shown in figure 1. Therefore, this project used only those variables which an investor can use to make a realistic data-driven decision. The filtered dataset contains only 28 variables from Lending club dataset as shown in Table 1.

Variables Name	Description	Variable Name	Description
Id	Loan ID	**grade**	Grade of the loan
loan_amnt	Loan Amount Approved	**emp_length**	Employment length of the borrower
funded_amnt	Loan Amount Funded	**home_ownership**	Home ownership of the borrower
term	Term of the loan	**annual_inc**	Annual Income of the borrower
int_rate	Interest Rate	**verification_status**	Verification status of loan
installment	Installments	**Issue_d**	Loan issue date
loan_status	Status of the loan	**purpose**	Borrower's purpose for loan
dti	Debt to income ratio	**delinq_2yrs**	No. of delinquencies in 2 years
earliest_cr_line	Date of borrower's account opening	**open_acc**	No. of opening accounts
pub_rec	Total derogatory public records	**fico_range_high**	Upper range of borrower fico score
fico_range_low	Lower range of borrower fico score	**revol_bal**	Borrower's revolving balance

revol_util	Revolving line utilization rate	**total_pymnt**	Total repays by borrower
last_pymnt_d	Borrower's Last payment date	**recoveries**	Amount recovered from borrower
addr_state	Borrower's resident state	**zip_code**	Borrower's resident zip code

Table 1. List of variables available to the investor

SOCIO-ECONOMIC FACTORS

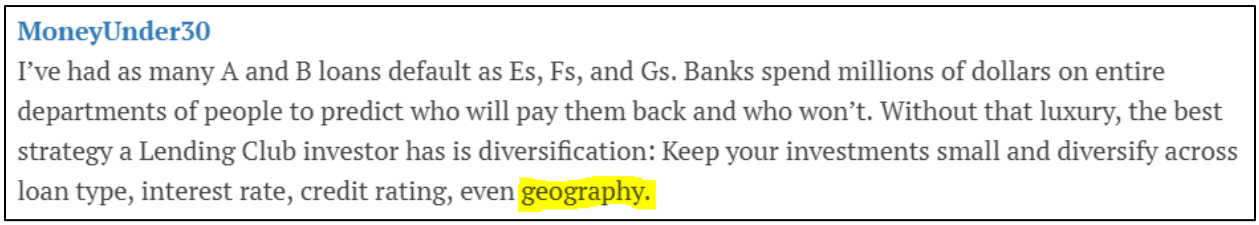

Figure 2. An investor's impression of using geographic factors to predict return

As highlighted by a personal financial advice website (Money Under 30), it is important to consider geographical information of a borrower while making an investment. As all the socio-economic attributes are specific to a geographical identity, incorporating these factors will help to understand the impact of location-specific social and economic indicators on loan default and rate of return prediction. Zip code in the Lending Club data contains the first three digits of the zip code for each loan id. Therefore, to utilize socio-economic features, attributes are population-weighted and aggregated for each three digits zip-code [3]. American Fact Finder provides the right platform to collect economic and social information for each zip code and for all the states. For our analysis, a set of 72 socio-economic factors are used and these belong to the following categories:

- **Employment status**
- **Commute to work**
- **Occupation**
- **Industry**
- **Class of worker**
- **Income and benefit**
- **Health insurance coverage**
- **Percentage of families below poverty level**

FEATURE ENGINEERING

A set of 13 new features were created based on loan issue_d, earliest_cr_line, dti, revol_bal, addr_state, fico_range_low, fico_range_high. Table 3 briefly describes how each of the features is generated. Some of the features with redundant information such as Installment & Interest Rate and loan & funded amount were removed. Also, features, derived by Lending Club, such as Grade was removed as the objective of this research is to use only the borrower information available at the time of investment. The final dataset after combining socio-economic indicators contains 118 features and about 355K borrowers.

New Features	Description
Cr_hist	Credit history of the borrower at the time loan was funded
revol_bal_wrt_loanamnt	Borrower's revolving amount with respect the funded amount
New_dti	Borrower's new dti based on repayment amount to the borrower's monthly income if the loan is approved [1]
state_	Top 10 state with highest Charged Off rates from 2007-2014: state_CA, state_NY, state_TX, state_FL, state_IL, state_NJ, state_PA, state_GA, state_OH, state_VA

Table 3. List of feature-engineered variables

ANNUALIZED RATE OF RETURN (ARR)

Lending Club does not provide ARR for each loan id in their historical data. Therefore, by studying their policy and investment strategies used by investors, ARR is calculated in the current dataset. ARR for a loan is calculated under the assumption that the amount received is immediately reinvested in a new loan entirely not in fraction and at a 3% prime rate, compounded monthly for 5 years [4]. This also undermines the time value of money, which restricts an investor to reinvest at a higher rate of interest. The formula to calculate the rate of return is given by

$$\frac{12}{T} * \frac{1}{f}\left\{\left[\frac{p}{m} * \frac{1-(1+i)^{m}}{1-(1+i)}\right] * (1+i)^{T-m} - f\right\}$$

f = Funded Amount
p = Total amount repaid and recovered
m = Number of months
p/m = Monthly payments that is re-invested
i = Prime rate
T = Term of the loan

In order to account for high default rate Lending Club charges higher interest rates for riskier loans, thereby providing low ARR for riskier loans as shown in figure 3. Surprisingly, ARR is for grade B loans is more than grade A loans and from grade B to G, ARR value drops sharply. However, there is no such precipitous drop for Lending Club's rate of return as we move from grade B to G. It is because Lending Club considers other factors such as annual charged off rate, potential losses on notes and service charges in their calculation. This research, for the sake of simplicity, will not consider any kind of deduction while calculating ARR and will use this value for further analysis. It is observed that, in the long run, ARR for Lending Club loans from grade A to G develops a bell curve pattern, which means ARR is higher in the middle and lower at both the ends. Since the study uses only 2015 historical loan data, the pattern may be arbitrary.

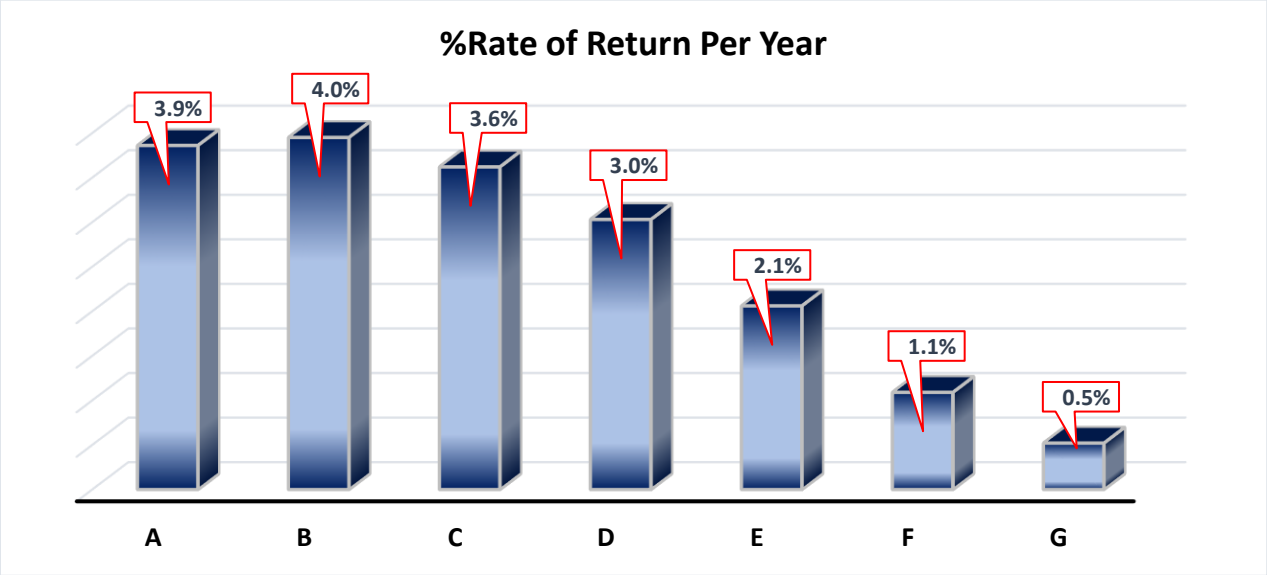

Figure 3. Comparison of annualized rate of return calculated for loans in 2015 from grades A-G

PRINCIPAL COMPONENT ANALYSIS

Statistical analysis with the substantial number of correlated variables leads to incorrect prediction. Hence, using Principal Component Analysis (PCA) dimension reduction technique, only the features preserving most of the information are retained. It is to trade a little accuracy for simplicity. As PCA is always performed on a symmetric correlation which means the matrix should be numeric and standardized, all the continuous features in the dataset were first standardized to unit variance and center to the mean to transform into a comparable scale and are segregated from the categorical variables. Python's sklearn preprocessing and decomposition package was used to perform standardization and PCA respectively. For this study, 10 Principal components are selected as they account for nearly 90% of the variance as shown in figure 4. Following is the python code to execute standardization and PCA

```
#Standardization of pca_data
#[X-avg(x)]/variance
from sklearn.preprocessing import scale
pca_scale= scale(pca_data)

#PCA and selection of 10 components
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
principalComponents = pca.fit_transform(pca.fit_transform(pca_scale))
```

Once the components were selected, categorical variables are combined to give a combined dataset of 10 continuous dimensionally reduced components and 16 categorical features to be used in further analysis

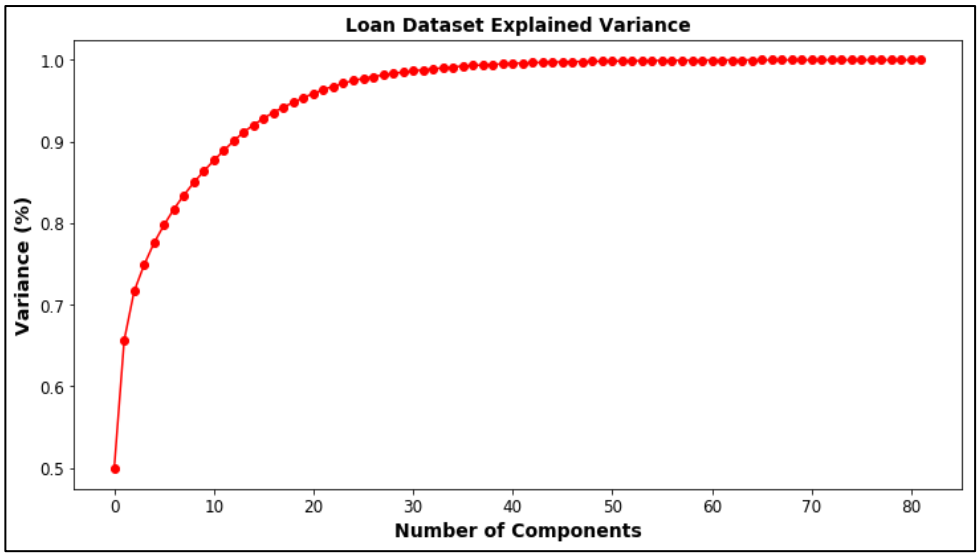

Figure 4. % Variance and principal components for Lending Cub loan dataset with socio-economic features

RESAMPLING

Lending club loan data has an imbalanced distribution of the target variable. About 80% of the loans have status Fully Paid whereas 20% have status Charged Off. Due to the disproportionate nature of the dataset, the model will have higher chances of predicting a non-event (Fully Paid) rather than an event (Charged Off). To counter this, 3 major resampling techniques are used.

- **Random under sampling (RUS)**
- **Random over sampling (ROS)**
- **Synthetic Minority Oversampling Technique (SMOTE)**

RUS will randomly pick a non-event and bring the proportion down to 50% whereas ROS will replicate randomly the minority classes and boost the proportion to 50%. A major drawback with RUS technique is the loss of valuable information whereas with ROS technique increasing likelihood of over-fitting. To avoid over-fitting and loss of information, SMOTE is used. SMOTE generates synthetic data using K-Nearest Neighbor and linear interpolation method. [5]. The resampling techniques are implemented using Sklearn.Imblearn Python Package and the following are the distribution.

	Charged Off	Fully Paid	Total Rows
Original dataset	71,232	286,671	357,903
Random Under Sampling	71,232	71,232	142,464
Random Over Sampling	286,671	286,671	573,342
SMOTE	286,671	286,671	573,342

Table 4. Total number of Charged Off and Fully Paid loans after resampling of dataset

Python imblearn package was used to implement resampling and the following are the codes to do so:

```
#Resampling
from collections import Counter
from imblearn.under_sampling import RandomUnderSampler
from imblearn.oversampling import RandomOverSampler
from imblearn.oversampling import SMOTENC

rus = RandomUnderSampler(random_state=42)
X_rus, Y_rus = rus.fit_resample(X, Y)

ros = RandomOverSampler(random_state=42)
X_ros, Y_ros = ros.fit_resample(X, Y)

smote=SMOTENC(categorical_features=[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])
X_sm, Y_sm = smote.fit_sample(X, Y)
```

In case of SMOTE, all the categorical variables are first dummy encoded and were explicitly mentioned in the function declaration.

CLASSIFICATION PROBLEM

The classification goal of this project is to correctly classify a loan status between Charged Off and Fully Paid using resampled dataset and predictive models such as Logistic Regression, neural network, Bayesian Network, Decision Tree, Lasso Regression, Gradient Boost, and Ensemble model. To perform an honest assessment, a separate scoring dataset containing 50 K borrowers from Lending Club loan history 2016 Q1 data and socio-economic factors for the year 2016 were preserved. The scoring dataset contains the same proportion of event and non-event as our current dataset.

These models are created using SAS® Enterprise Miner™ nodes such as data partition, model comparison, ensemble, and other models. Since the original sample was resampled to balance the proportion, it is important to regulate the prior probabilities in the process otherwise posterior probabilities will be incorrect and so will be the classifications. To let SAS® Enterprise Miner™ know about the prior distribution in the population, prior probabilities were assigned in the decision processing option in the input source node as shown in figure 5. Dataset was then further divided into a train and test in the ratio of 70/30, which a standard practice in analytics. Once the models are trained and validated, performance metrics such as KS Statistics, ROC Index and Sensitivity are used to evaluate them and find the champion model.

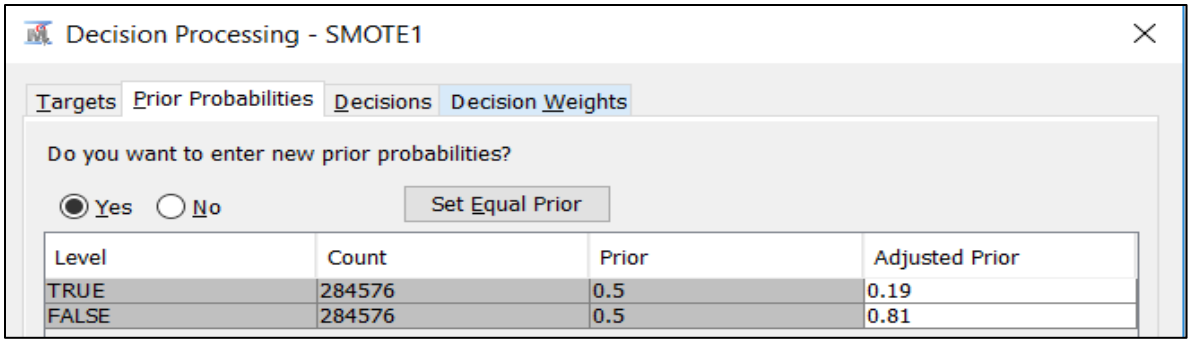

Figure 5. Prior probability setting in the input data source node in SAS® Enterprise Miner™

In SAS® Enterprise Miner™, the default decision cutoff is 0.5, which is more than 19% Charged Off rate in the population so it is necessary to choose the right cutoff value for our analysis. To find the right cutoff, either the model must emphasize on sensitivity value or the precision in the event prediction. Since the objective of the study is to find a set of non-default or Fully Paid loans for investors to invest money, so it is required to reduce the number of false negative as it can cost investors the real money. The probability cutoff should be close to the event probability in the population [6], therefore, for our analysis, cutoff value of 0.2 is considered to calculate sensitivity, overall classification rate for the champion model. The entire Enterprise Miner process flow diagram for the one resampled dataset is show in figure 6.

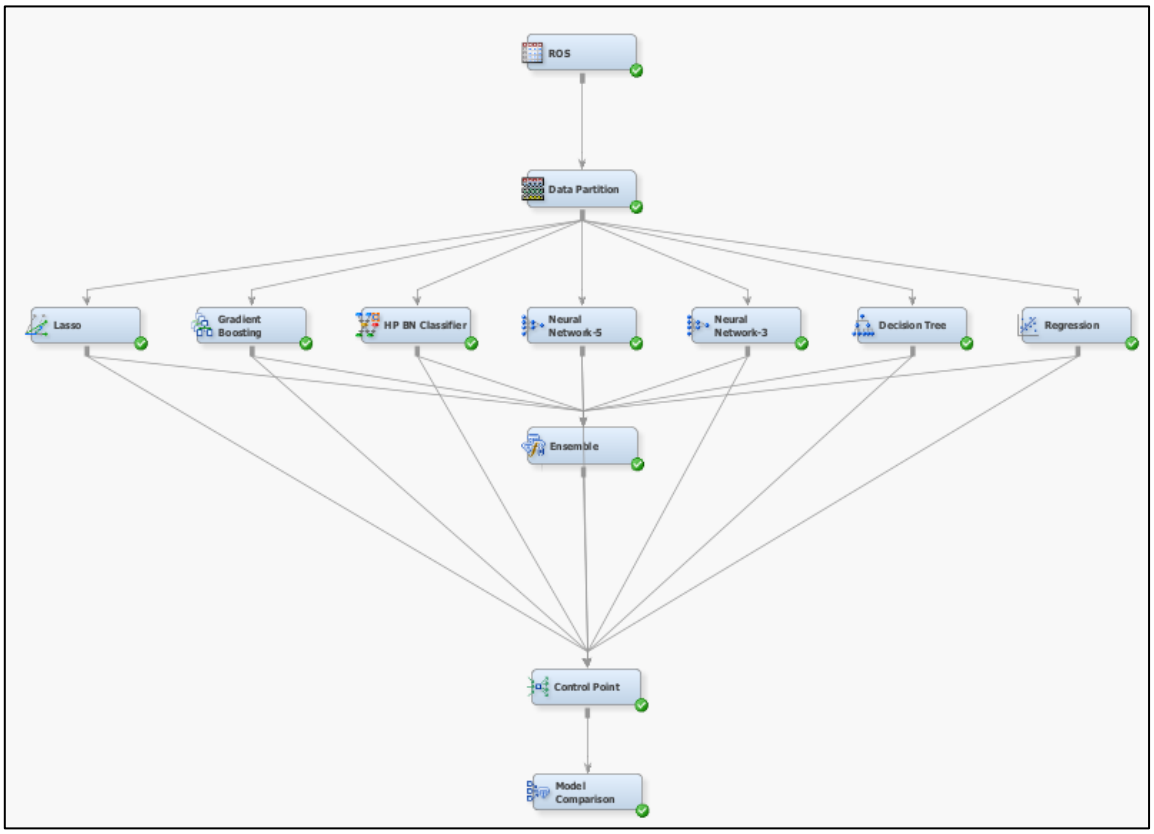

Figure 6. SAS® Enterprise Miner™ process flow diagram for classification modeling

PERFORMANCE STATISTICS

To compare the performance statistics of the models, SAS® Enterprise Miner™ has model comparison node, which can compare all the models with each other on a desired metrics. For our analysis, model comparison is conducted using validation misclassification rate. From the model comparison node, ROC Index, KS statistics and Sensitivity of all the models on different resampled datasets are presented in the table 5. As highlighted in the table, Neural Network model with 3 hidden units, trained and validated on SMOTE sampling dataset, is our champion classification model as it has the highest sensitivity value of 46.6%. From figure 7 it is clear how models are performing in terms of capturing true positives from the validation dataset.

	ROC Index			KS Statistics			Misclassification Rate			Sensitivity		
	ROS	RUS	SMOTE	ROS	RUS	SMOTE	ROS	RUS	SMOTE	ROS	RUS	SMOTE
Logistic Regression	70%	70%	79%	0.29	0.28	0.54	48%	48%	39%	5.6%	6.0%	41.7%
Lasso Regression	70%	70%	79%	0.29	0.28	0.54	48%	48%	39%	5.6%	5.6%	41.6%
Decision Tree	64%	66%	64%	0.26	0.26	0.3	46%	48%	50%	6.1%	9.4%	33.4%
Gradient Boost	50%	60%	50%	0	0.14	0	50%	50%	50%	0.0%	0.0%	0.0%
Bayes Network	71%	68%	77%	0.27	0.27	0.51	50%	47%	42%	6.6%	8.4%	35.2%
Neural Net (3)	70%	71%	**79%**	0.29	0.29	**0.54**	48%	47%	**29%**	5.6%	6.6%	**46.6%**

Neural Net (5)	71%	71%	79%	0.29	0.29	0.55	48%	47%	38%	6.6%	7.7%	45.9%
Ensemble	70%	71%	78%	0.29	0.29	0.54	48%	48%	42%	4.7%	5.0%	41.5%

Table 5. Performance statistics of all the model tested on resampled data

Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Reg2	Regression	TRAIN	outcome		116975	192464	8205	83694
Reg2	Regression	VALIDATE	outcome		50129	82452	3550	35873
Boost2	Gradient Boosting	TRAIN	outcome		200669	200669	0	0
Boost2	Gradient Boosting	VALIDATE	outcome		86002	86002	0	0
Neural3	Neural Network-3	TRAIN	outcome		107192	190655	10014	93477
Neural3	Neural Network-3	VALIDATE	outcome		45905	81712	4290	40097
Tree2	Decision Tree	TRAIN	outcome		133873	185560	15109	66796
Tree2	Decision Tree	VALIDATE	outcome		57265	79445	6557	28737
Ensmbl2	Ensemble	TRAIN	outcome		117480	193230	7439	83189
Ensmbl2	Ensemble	VALIDATE	outcome		50292	82750	3252	35710
HPBNC2	HP BN Classifier	TRAIN	outcome		129949	193681	6988	70720
HPBNC2	HP BN Classifier	VALIDATE	outcome		55654	82943	3059	30348
LARS2	Lasso	TRAIN	outcome		117071	192354	8315	83598
LARS2	Lasso	VALIDATE	outcome		50166	82386	3616	35836
Neural4	Neural Network-5	TRAIN	outcome		108418	191454	9215	92251
Neural4	Neural Network-5	VALIDATE	outcome		46454	81981	4021	39548

Figure 7. Event classification table output for random under sampling dataset

To see how the same set of models perform on the original data, a sample set of 100K records is obtained having 81%/19% proportion of fully paid and charged off loans respectively. It is observed that no model is performing better than the champion model as the misclassification rate is close to 19%, which means, the model is only able to classify non-events or fully paid loans and almost misclassifying the entire event or charged off loans. With respect to the original dataset, SMOTE sampling method has done better.

CUTOFF NODE & SCORING

As discussed earlier, the cutoff for classification into fully paid or charged off is 0.5 even after the prior probability adjustment. Therefore, it is important to change the cutoff close to 0.19 and estimate the sensitivity value of the model. For our analysis, the cutoff was set at 0.2. This was achieved using SAS® Enterprise Miner™ Cutoff node. It allows changing the cutoff value according to different criteria such as precision, sensitivity, accuracy, etc. In the Cutoff node user-input option was selected and 0.2 value was added. As shown in figure 8, the true positive rate and overall classification rate increased to 73% and 77% respectively at the desired cutoff. With the increase in true positive rate, false negative rate for the model has gone down significantly, which prevents an investor to lose money by investing on misclassified charged off loans. Once the cutoff was decided, to further test the model, an honest assessment was done on scoring dataset using SAS code and Score node [7] as shown in figure 9. The sensitivity on the scoring dataset is 42.8% which our model at 0.2 cutoff value is the best model without much difference in the training and test performance.

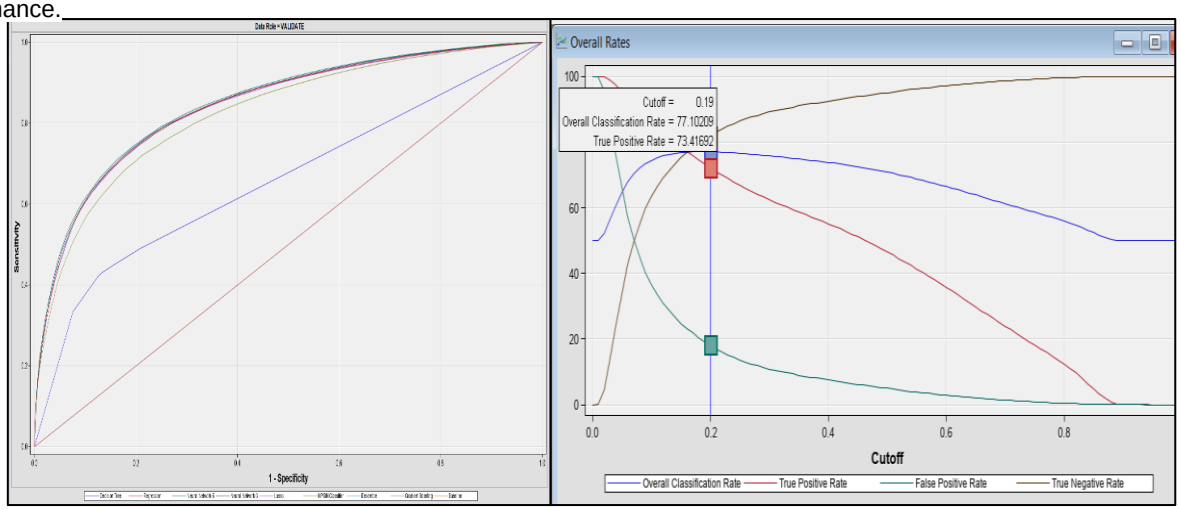

Figure 8. Overall classification rate, true positive rate and true negative rate at cutoff =0.2

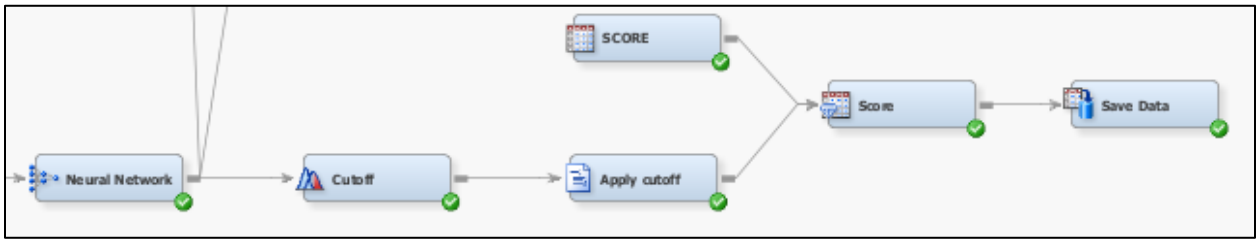

Figure 9. SAS® Enterprise Miner™ process flow diagram for honest assessment using score data

REGRESSION PROBLEM

Loans identified are then moved to stage 2. Stage 2 of the project is to calculate expected ARR for loans. For the regression objective, about 200K borrowers were randomly selected from the original dataset and converted into SAS dataset using SAS® Enterprise Guide™. Multiple regressive models, such as Linear regression, Neural Network with 3 and 5 hidden units, decision tree, Gradient Boosting, and Ensemble, were then built to find the best performing model based on average squared error.

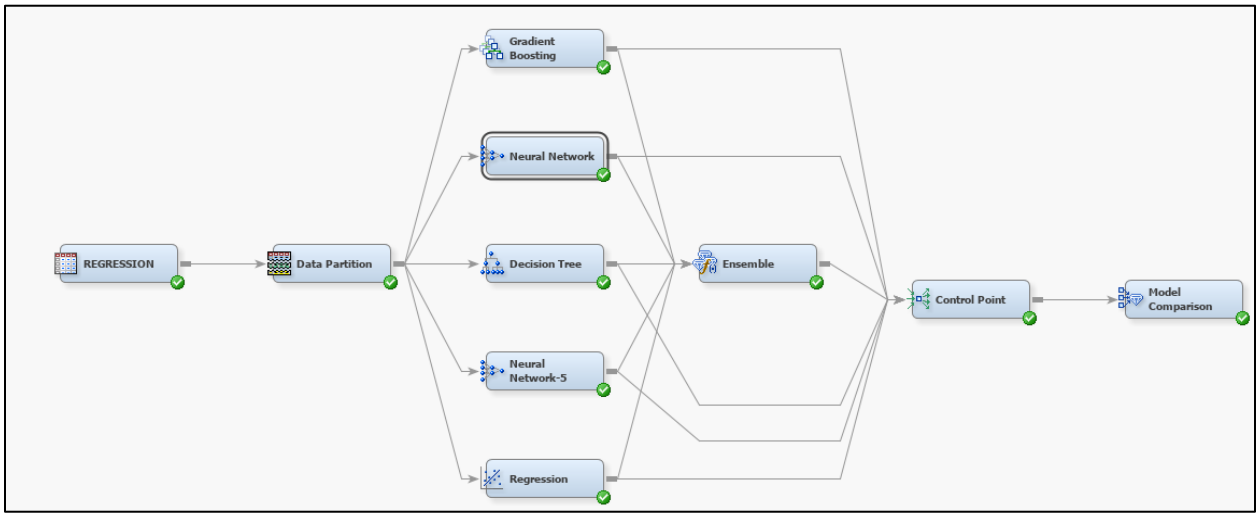

Figure 10. SAS® Enterprise Miner™ process flow diagram for Annualized Rate of Return (ARR) prediction

As shown in figure 10, models were tested on a validation dataset containing 30% of the dataset. Like the classification problem, a model comparison node was used to compare the performance of the model and select the champion model based on validation average squared error. Since there is no need for any resampling and balancing of classes for regression, separate scoring method is not required. On evaluating the average squared error value as shown in figure 11, again Neural Network model with 3 hidden units is the champion regression model as it has the lowest average squared error compared to other models. Neural Network model is further used for strategizing investment portfolio.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	T	Selection Criterion: Valid: Average Squared Error
Y	Neural	Neural	Neural Network	return		0.003354
	Neural2	Neural2	Neural Network-5	return		0.003355
	Ensmbl	Ensmbl	Ensemble	return		0.003361
	Tree	Tree	Decision Tree	return		0.003369
	Reg	Reg	Regression	return		0.003376
	Boost	Boost	Gradient Boosting	return		0.003466

Figure 11. SAS® Enterprise Miner™ model comparison node output for regression models

EXPECTED RETURN & INVESTMENT PORTFOLIO

Expected return on an investment is the expected value of the probability distribution of all the possible outcomes of a loan. For example, if the investor invested on a charged off loan, the return will be different if the same loan is fully paid by the borrower. [5] Therefore, the expected return will consider all the possible outcomes and their chances of occurring. The formula to calculate expected return is given by:

$$\boldsymbol{Expected\ Return = P_{ChargedOff} * Return_{ChargedOff} + P_{FullyPaid} * Return_{FullyPaid}}$$

To calculate the expected return, two separate datasets containing charged off and fully paid loans are trained using the champion regression model. ARR value for each of the models is predicted on the scoring dataset, which was previously used in the classification problem. In conjunction with SAS® Enterprise Miner™ Save Data node, score predictions were saved to a local drive for further analysis

Figure 12 illustrates the SAS® Enterprise Miner™ process flow diagram of training two additional models to predict the expected return in two different scenarios. Once the respective returns are predicted, the values are then sorted on a common variable and expected return was calculated in Microsoft Excel using the above formula as shown in figure 13.

The objective of this study was to find the optimal number of loans an investor should invest to get the best return. This will help to understand the volatility in the rate of return when investments are done in large number and model's scalability. This will also help to understand how diversification of loans will stabilize the rate of return in the long run. This can be easily visualized using a plot between average rates of return and an increasing number of loans as shown in figure 13. For the 50,000 loans in the score data, the average ARR is 2.44%. This value is lower than what can be generally expected because of loss of any information or features that this study didn't consider. However, 2.44% can be a realistic return in this scenario as Lending Club return is between 3-8%.

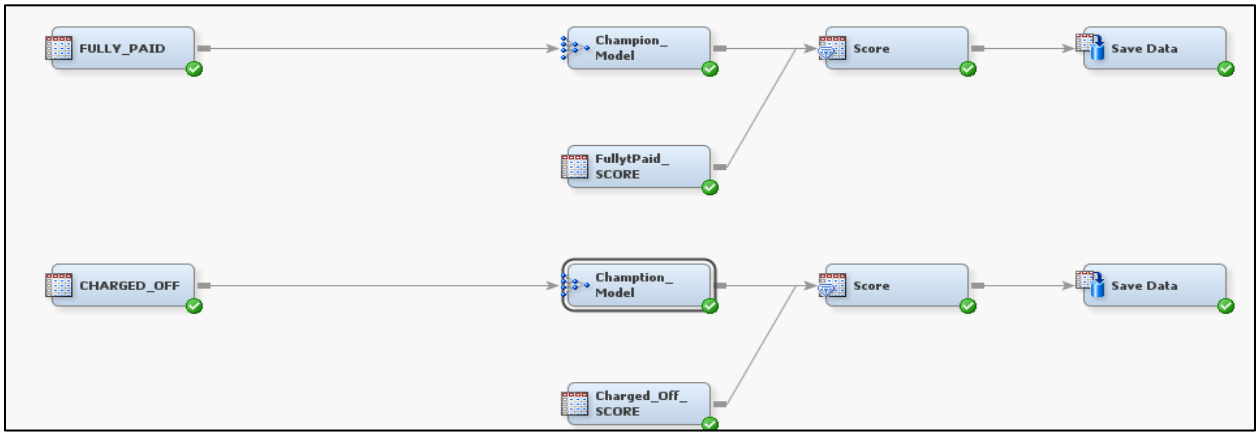

Figure 12 SAS® Enterprise Miner™ process flow diagram for training champion model on Fully Paid and Charged Off dataset

As illustrated in figure 13, average ARR value decreases with increasing portfolio size. The reason for this kind of trend is the decreasing availability of good loans or loans with good returns.

Charged Off Probability	Charged Off Return	Fully Paid Probability	Fully Paid Return	Expected_Return
0.017102912	-0.048230824	0.982897088	0.084755839	0.08248138
0.048548783	-0.048429255	0.951451217	0.085190723	0.078703636
0.043069623	-0.053390853	0.956930377	0.08199761	0.07616648
0.044363676	-0.076978195	0.955636324	0.082212392	0.075150112
0.054537914	-0.072597592	0.945462086	0.082467185	0.074010275
0.044521805	-0.075395059	0.955478195	0.080534051	0.073591806
0.089449436	-0.062206402	0.910550564	0.086456727	0.073158894
0.087269497	-0.063991768	0.912730503	0.085892844	0.072812489
0.054415083	-0.07862639	0.945584917	0.08120976	0.072512263

Figure 13. Expected Return calculation in Excel (showing top 10 rows)

Also, Lending Club has demonstrated how with diversification-spreading an investment equally across hundreds of loans can derive solid return.

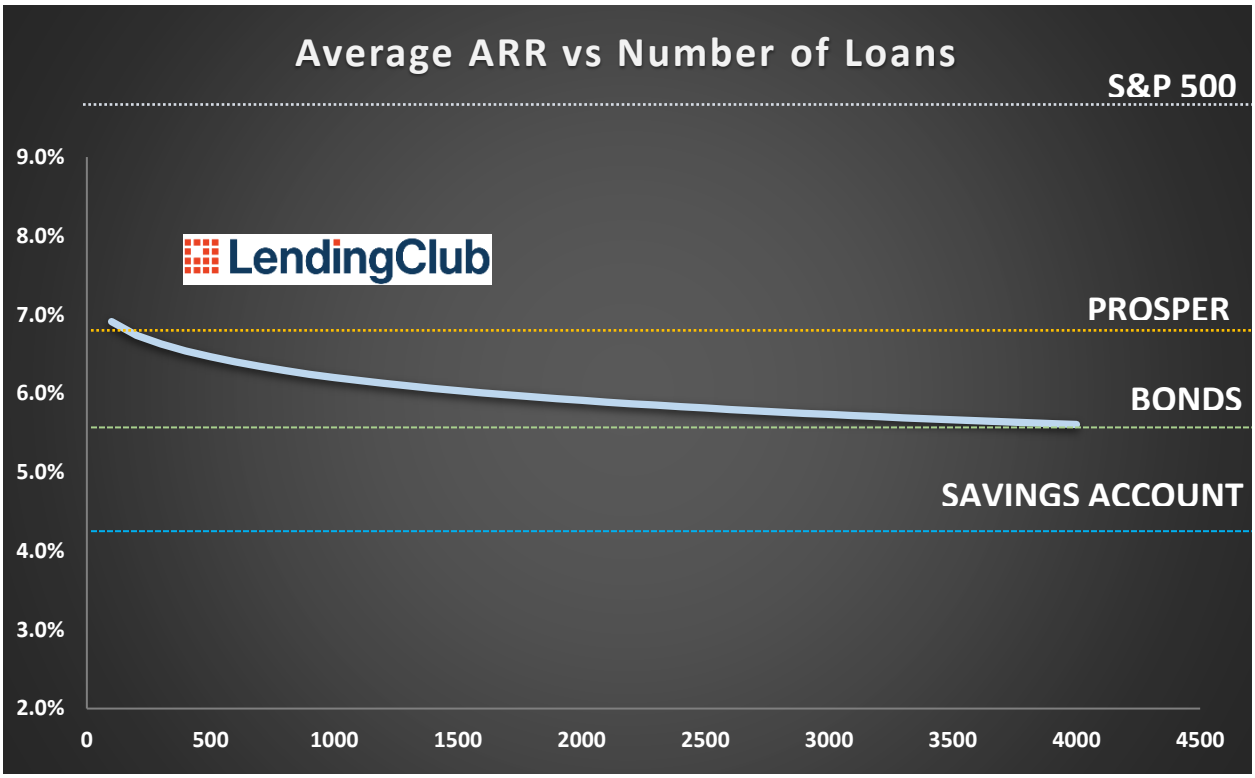

Figure 14. Comparative plots of average annualized rate of return of different types of investments

In order to understand how Lending Club returns are crucial, it is important to compare the rate of return per year with other types of investment options such as S&P 500, Prosper lending, Bonds & Savings Account. From the figure 14, Lending Club return on investment are higher than that of Prosper, Bonds & Savings accounts. However, investors must invest in at least 1,500 loans to have higher return as compared to the returns from Prosper Loans. This comparison is done for investments made for a period of 5 years only. With a greater number of years, rate of return could be much higher, and this analysis mayn't remain valid.

SUMMARY

The research has some interesting finding both in stage I and stage II of the project. In the stage I, SMOTE balancing technique has yield the highest sensitivity rate of 46.6% at 0.5 as the decision cutoff as compared to Random under sampling and Random over sampling techniques. When the decision cutoff was changed to 0.2, sensitivity value increased to 73% whereas false negative rate decreased to 27%. In stage II, the average expected annualized rate of return for 50,000 loans is 2.44% which is slightly lower than the Lending Club' s range of return per year. When compared to four other types of investments such as S&P 500, savings account, bonds, and Prosper Lending, rate of return for Lending Club is better than that of saving's account, bonds and Prosper.

REFERENCE

[1] Anahita Namvar. 2018. "Credit risk prediction in an imbalanced social lending environment". https://arxiv.org/abs/1805.00801

[2] Peiqian Li. "Lending Club Loan Default and Probability Prediction". https://cs229.standard.edu/proj2018/report/69/pdf

[3] Shunpo Chang. 2015. "Predicting Default Risk of Lending Club Loans". https://www.semanticscholar.org/paper/Predicting-Default-Risk-of-Lending-Club-Loans-S-Chang-Kim/6f64741e33b82e2dea0fe1179678e14bae05555b

[4] Maxime.C.Cohen, 2018. “Data-Driven Investment Strategies for Peer-to-Peer Lending: A Case Study for Teaching Data Science”. https://www.liebertpub.com/doi/full/10.1089/big.2018.0092

[5] 2019. “Predict Lending Club ‘s Loan Data”. https://rstudio-pubs-static.s3.amazonaws.com/203258_d20c1a34bc094151a0a1e4f4180c5f6f.html

[6] Yogen Shah. 2012. “Use of Cutoff and SAS Code Nodes in SAS® Enterprise Miner™ to Determine Appropriate Probability Cutoff Point for Decision Making with Binary Target Models” https://support.sas.com/resources/papers/proceedings12/127-2012.pdf

[7] Shivateja Reddy Kandula. 2018. “Predicting the risk of attrition for undergraduate Students using SAS® Enterprise Miner™. https://analytics.ncsu.edu/sesug/2018/SESUG2018_Paper-243_Final_PDF.pdf

ACKNOWLEDGMENTS

Thanks to Information Research and Information Management of Oklahoma State University, Stillwater for allowing me to work on this data and providing timely inputs. Thanks to my Professors Dr. Goutam Chakraborty and Dr. Miriam McGaugh for tutoring the concepts of Data Mining and Machine Learning.

CONTACT INFORMATION

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

Priyabrata Thatoi
Graduate Student of Business Analytics (Class of 2020)
Oklahoma State University, Stillwater
+1 (405)762-6874
Priyabrata.Thatoi@okstate.edu
https://www.linkedin.com/in/pthatoiosu/