LOAN DEFAULT PREDICTION

Amarjeet Cheema, Oklahoma State University

Miriam McGaugh, PhD, Oklahoma State University

ABSTRACT

Interest on loans and associated fees are some of the biggest revenue sources for most banks and credit unions. More than 44 million borrowers collectively owe about $3.5 trillion in total outstanding consumer credit

As of October 2015. However, more than 1 million people default on loans each year. A report from The Urban Institute, a non-profit research institute, found that nearly 40% of borrowers are expected to default on their student loans by 2023.

Considering the magnitude of risk and financial loss involved, it is essential for banks to give loans to credible applicants who are highly likely to pay back the loan amount. The objective of my project is to assess the likelihood of loan default based on customer demographics and financial data. Furthermore, as an outcome of this project we found the most significant variables that contribute to determining loan default. The project will also categorize loan type based on highest risk of default. Such insights will help the banks in significantly reducing the risk of losing money associated with loans. The original dataset had 887,000 observations and 24 columns; however, it was filtered to 208,941 rows of data for which the loan status was either fully paid or default. The resulting dataset had 209,000 rows with 24 variables. SAS Enterprise Miner was used to create logistic regression and decision trees models with different configurations and SAS Viya was used for data visualization.

INTRODUCTION

Banks are essential bodies of the society which not only serve as a secure vault for storing money but also help people in time of need by providing loans and credit. Banks provide loans based on credit history and credibility of the applicant. However, the rate of loan default is increasing exponentially. According to Forbes, say that nearly 40 percent of borrowers are expected to default on their student loans by 2023. Considering the statistics, instead of making money from loan interest, banks will suffer a huge capital loss. In order the prevent the loss, it is very important to have a system in place which will accurately predict the loan defaulters even before approving the loan.

DATA DESCRIPTION

- Data is in form of excel worksheet
- Worksheet contains 887K observations and 24 columns
- Data consists of customer demographics as well and financial details such as total amount funded, every month installment (EMI) and rate of interest
- Data also has housing and customer employment information such as housing ownership, years in job and annual income
- Information regarding delinquency and late repayment is also present in the data.
- Data set was extracted from Kaggle

Variable	Data type	Description
id	ID	Unique Identification of the borrower
loan_amnt	Numeric	Loan amount applied by the borrower
funded_amnt_inv	Numeric	Actual amount approved for the borrower
term	Numeric	Number of months for loan repayment
int_rate	Numeric	Annual Interest rate on loan
installment	Numeric	Monthly installment the borrower has to pay
grade	Categorical	Category of loan
sub_grade	Categorical	Subcategory of the loan
emp_length	Numeric	Duration of employment of the borrower
home_ownership	Categorical	To check if the borrower own a home or stays in rent
annual_inc	Numeric	Annual income
verification_status	Categorical	Income verification status
loan_status	Categorical	Indicator to show if borrower has fully paid the loan or has defaulted
zip_code	Numeric	Home address zip code of the borrower
addr_state	Numeric	Home address state of the borrower
dti	Numeric	Debt to income ratio
delinq_2yrs	Numeric	Indicator to test if borrower has any delinquency record in last two years
total_acc	Numeric	Total number of accounts of the borrower
total_pymnt_inv	Numeric	Total amount repaid by the borrower
total_rec_late_fee	Numeric	Total late fee paid by the borrower
emp_title	Categorical	Company of the borrower
title	Categorical	Purpose of taking the loan
purpose	Categorical	Purpose of taking the loan
desc	Text	Detailed purpose of the loan

Figure 1. Data Dictionary

DATA PREPARATION

The original dataset had 887K observations and 24 columns but some of the observations were not related to our business goal. Rows of data were filtered for loan status as either fully paid or default. The resulting dataset had 209k rows with 24 variables.

Variable Levels Summary		
Variable	**Role**	**Frequency**
id	ID	208941
loan_status	TARGET	2

Figure 2. Data Type Classification

DATA EXPLORATION

Initial exploratory data analysis was performed using the StatExplore node to understand the variation and range of the variables.

Interval Variable Summary Statistics									
Data Role = Train									
Variable	Role	Mean	Standard Deviation	Non-Missing	Missing	Minimum	Maximum	Skewness	Kurtosis
annual_inc	INPUT	74118.78	59048.26	208941	0	3000	64000	33.64	3066.63
delind_2yrs	INPUT	0.25	0.73	208941	0	0	0	5.9	68.7
dti	INPUT	16.16	7.7	208941	0	0	15.77	0.24	-0.46
installment	INPUT	413.4	244.18	208941	0	15.69	360.38	1.03	0.97
int_rate	INPUT	13.29	4.27	208941	0	5.32	13.11	0.4	-0.17
loan_amt	INPUT	13357.17	8060.13	208941	0	500	12000	0.87	0.19
total_pymnt_inv	INPUT	15028.05	9464.19	208941	0	0	12817.68	0.99	0.65

Class Variable Summary Statistics							
Data Role=Train							
Variable_Name	Role	No of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
addr_state	INPUT	51	0	CA	17.22	NY	8.29
exp_length	INPUT	12	0	10+ years	30.69	2 years	9.39
grade	INPUT	7	0	B	31.94	C	25.38
home_ownership	INPUT	6	0	Mortgage	50.48	RENT	40.8
purpose	INPUT	14	0	debt_consol	58.18	credit_card	20.33
term	INPUT	2	0	36 months	80.54	60 months	19.46
verification_status	INPUT	3	0	Not verified	35.49	Verified	35.45
loan_status	INPUT	2	0	Fully Paid	99.42	Default	0.58

Figure 1. Summary Statistics

DATA SAMPLING

The data set was highly unbalanced having 99.42% observations corresponding to loan status as ‘fully paid’. On the other hand, just 0.58% observations corresponding to loan status as ‘default’. To build a good prediction model on this data, equal sampling was performed on the data set to balance the data and make it bias free.

Summary Statistics for Class Targets				
Data=DATA				
Variable	Numeric Value	Formatted Value	Frequency	Percent
loan_status	.	Default	1219	0.59
loan_status	.	Fully Paid	207722	99.42
Summary Statistics for Class Targets				
Data=SAMPLE				
Variable	Numeric Value	Formatted Value	Frequency	Percent
loan_status	.	Default	1219	50
loan_status	.	Fully Paid	1219	50

Figure 4. Data Glimpse Before and After Sampling

DATA SPLITTING

In order to avoid overfitting the data, the Partition node was used to divide the entire sample data into 70/30 ratio with 70% training data and 30% validation data. Below is a glimpse of the data before and after splitting.

Data=TRAIN				
Variable	Numeric Value	Formatted Value	Frequency Count	Percent
loan_status	.	Default	853	50.03
loan_status	.	Fully Paid	852	49.97
Data=VALIDATE				
Variable	Numeric Value	Formatted Value	Frequency Count	Percent
loan_status	.	Default	366	49.94
loan_status	.	Fully Paid	367	50.06

Figure 5. Loan Status Frequency in Training and Validation data

OUTLIER HANDLING

Sometimes extreme values known as outliers change the parameters of the model and highly influence the performance of the model. Out of the total 1,705 observations, 1,348 observations were outside 3 standard deviation of mean therefore those observations were removed.

To deal with these extreme values we used filter node and eliminated such observations.

Number of Observations			
Data Role	Filtered	Excluded	Data
Train	1557	148	1705

Figure 6. Number of Filtered and Excluded Observations

IMPUTATION

Missing values were significantly reducing the training data set, therefore missing data values were imputed. The missing interval values were imputed using mean and the missing categorical values were imputed by the mode value.

Variable delinq_2yrs had missing values which were replaced by mean value of 0.13 and dti had minimum missing values which were replaced by 18.3

Variable Name	Impute Method	Imputed Variable	Impute Value	No of Missing for Train
REP_annual_inc	MEAN	IMP_REP_annual_inc	65073.37	10
REP_delinq_2yrs	MEAN	IMP_REP_delinq_2yrs	0.13	49
REP_dti	MEAN	IMP_REP_dti	18.3	3
REP_installment	MEAN	IMP_REP_installment	416.73	11
REP_total_acc	MEAN	IMP_REP_total_acc	24.52	15
REP_total_pymnt_inv	MEAN	IMP_REP_total_pymnt_inv	10089.44	6

Figure 7. Missing Value Imputation

DATA MODELLING

After data preparation, the data was ready for modelling. Figure 8 shows the modeling diagram that was used for this analysis. Decision tree and regression were the two statistical models utilized in the model.

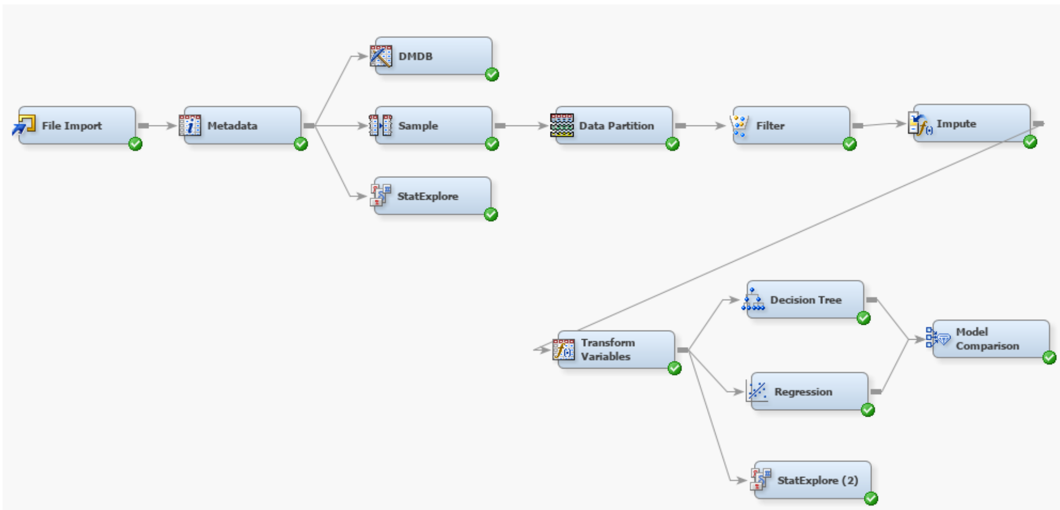

Figure 8. Model Diagram

Decision Tree:

A decision tree node was used to build the model keeping loan status as the target variable and event as 'Default'. Below is the list of variables used for decision tree modeling.

Name	Use	Report	Role	Level
id		No	ID	Nominal
dataobs		No	ID	Interval
TG_purpose	Default	No	Input	Nominal
TG_addr_state	Default	No	Input	Nominal
term	Default	No	Input	Nominal
home_ownership	Default	No	Input	Nominal
SQRT_int_rate	Default	No	Input	Interval
TG_emp_length	Default	No	Input	Nominal
TG_grade	Default	No	Input	Nominal
LOG_delinq_2yrs	No	No	Input	Interval
verification_status	Default	No	Input	Nominal
PWR_total_pymnt_inv	Default	No	Input	Interval
PWR_annual_inc	Default	No	Input	Interval
SQRT_dti	Default	No	Input	Interval
SQRT_loan_amnt	Default	No	Input	Interval
SQRT_installment	Default	No	Input	Interval
loan_status	Yes	No	Target	Nominal

Figure 9. Model Input and Target Variables

Logistic Regression Model:
The logistic regression model with link function as logit was used to analyze the data. The variable selection model was kept as stepwise and validation misclassification rate was used as selection criteria.

Model comparison node:
To pick the best model based on the performance on the validation data, a model comparison node was used keeping misclassification rate as the decision parameter.

RESULTS

After running the model comparison node to compare the performance of the decision tree and the logistic regression model on the dataset, the logistic regression model was selected as the final model based on validation misclassification rate, which was 4.4%. This rate was less than other models and it also had high sensitivity of 95.63% and very high specificity as 97.55%.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Reg	Reg	Regression	loan_status	loan_status	0.043656
	Tree	Tree	Decision Tree	loan_status	loan_status	0.065484

Figure 10. Model compassion results

Event Classification Table							
Model selection based on Valid: Misclassification Rate							
Model	Description	Data Role	Target	False Negative	True Negative	False Positive	True Positive
Reg	Logistic Regression	Train	loan_status	55	758	10	734
Reg	Logistic Regression	Validate	loan_status	23	358	9	343
Tree	Decision Tree	Train	loan_status	55	755	13	734
Tree	Decision Tree	Validate	loan_status	33	352	15	333

Figure 11. Classification table for logistic regression model

Analysis of Maximum Likelihood Estimates						
Parameter	**loan_status**	**DF**	**Estimates**	**Standard Error**	**Wald Chi-Square**	**Pr>ChiSq**
Intercept	Default	1	13.23	1.97	44.98	<.0001
PWR_annual_inc	Default	1	-3.48	1.29	7.24	0.0071
PWR_total_pymnt_inv	Default	1	-46.31	2.9	254.11	<.0001
SQRT_installment	Default	1	29.81	2.02	217.43	<.0001
SQRT_int_rate	Default	1	9.8	2.44	16.11	<.0001
TG_grade A	Default	1	2.63	0.95	7.63	0.0057
TG_grade B	Default	1	1.45	0.41	12.17	0.0005
TG_grade C	Default	1	0.62	0.25	6.08	0.0137
TG_grade D	Default	1	-0.48	0.33	2.14	0.1431
TG_grade E	Default	1	-2	0.57	12.41	0.0004
term 36 months	Default	1	-1.6	0.22	55.08	<.0001

Figure 12. Parameter estimate tables

- A 1-unit increase in PWR_annual_inc, decreases the odds of loan_status being default by 0.031.
- A 1-unit increase in PWR_total_pymnt_inv, decreases the odds of loan_status being default by less than 0.001.
- A 1-unit increase in SQRT_installment, decreases the odds of loan_status being default by 999.
- A 1-unit increase in SQRT_int_rate, decreases the odds of loan_status being default by 999.
- For TG_grade A the odds of loan_status being default is 127.112 times the odds of Other.
- For TG_grade B the odds of loan_status being default is 39.48 times the odds of Other.
- For TG_grade C the odds of loan_status being default is 17.04 times the odds of Other.
- For TG_grade D the odds of loan_status being default is 5.668 times the odds of Other.
- For TG_grade E the odds of loan_status being default is 1.24 times the odds of Other.

Odds Ratio Estimates			
Effect		**loan_status**	**Point Estimate**
PWR_annual_inc		Default	0.031
PWR_total_pymnt_inv		Default	<0.001
SQRT_installment		Default	999
SQRT_int_rate		Default	999
TG_grade A	vs Other	Default	127.112
TG_grade B	vs Other	Default	39.48
TG_grade C	vs Other	Default	17.04
TG_grade D	vs Other	Default	5.668
TG_grade E	vs Other	Default	1.24
term 36 months	vs 60 month	Default	0.04

Figure 13. Odds ratio table

CONCLUSION

Based on the results, it can be concluded that this model can predict loan defaulters with an accuracy of 95.63%. Companies can employ similar models and potentially avoid giving loans to applicants who are highly likely to default. Doing so will reduce risk and financial loss for lending companies. As with all predictive models, data should be monitored and re-evaluated on a regular basis.

REFERENCES

1. 40% of the borrowers may default on their student loans by Zack Friedman, Posted on Nov 5, 2018, https://www.forbes.com/sites/zackfriedman/2018/11/05/student-loans-default-strategy/#e2d5c6f17c69
2. A look on the shocking student loan debt, Posted on Feb 4,2019, https://studentloanhero.com/student-loan-debt-statistics/

   https://www.cometfi.com/student-loan-debt-statistics
3. Car loan Statistics, https://www.finder.com/car-loan-statistics
4. More than 1 million people default on their student loan every year on Aug 13, 2018, https://www.cnbc.com/2018/08/13/twenty-two-percent-of-student-loan-borrowers-fall-into-default.html

CONTACT INFORMATION

Amarjeet Singh Cheema

Oklahoma State University

amarjeet.cheema@okstate.edu

+1 (405)-385-2876