Predict Loan Defaulter

MIS 6324.003 : Business Analytics With SAS

UNIVERSITY OF TEXAS AT DALLAS

November 13, 2017

Authored by: Shivam Parashar and Apoorva Purohit (Group 5)

EXECUTIVE SUMMARY

Objective

To develop predictive model using the loan dataset to predict whether a customer is a defaulter based on different attributes from the dataset.

Target Audience

The banks would be interested in buying our model to filter out the potential defaulters.

Competition

There are many models for this type of prediction but given a real time data set we can build a model which would have an exponential growth opportunity.

Risk

There is risk that model may not perform as per desired expectation with a different data set. We may end up providing a loan to a potential defaulter or fail to grant loan to a good customer.

Conclusions

Our model has a low misclassification rate in training and validation datasets. We can identify the potential defaulters efficiently.

Project Motivation

**(i)= is the reference number.

The number of defaulted federal student loans hit a new high in 2016: about 8 million borrowers have given up paying on more than $137 billion in education debts (1).

That means at least one out of every six people who have any federal student debt haven't made a payment on their loans for at least nine months, says Jessica Thompson, research director for The Institute for College Access and Success.

There is also a similar trend observed across different domains.

This motivated us in developing a model which would help the loan provider to give loan to a customer who is potentially more accomplished to pay up the loan back. This would help them increase their profit and reduce the defaulter to a certain extent.

Data Description

Source of Dataset: Kaggle.com (second hand dataset)(2)

Aim:

- To analyze the loan defaulters across a range of factors of age, gender, education status and principal amount borrowed by the customer.
- To develop predictive model using the loan dataset to predict weather a customer is a defaulter based on different attributes from the dataset.
- To do a comparative study between different algorithms such as Logistic Regression, Bayesian Networks classifiers, Random Forest and decision trees.

Dataset:

- This data set includes customers who have paid off their loans, who have been past due and put into collection without paying back their loan and interests, and who have paid off only after they were put in collection.
- The financial product is a bullet loan that customers should pay off all of their loan debt in just one time by the end of the term, instead of an installment schedule. Of course, they could pay off earlier than their pay schedule.

Variables	***Description***
Loan id	A unique loan number assigned to each loan customers
Loan status	Whether a loan is paid off, in collection, new customer yet to payoff, or paid off after the collection efforts
Principal	Basic principal loan amount at the origination
Terms	weekly (7 days), biweekly, and monthly payoff schedule
Effective date	When the loan got originated and took effects
Due date	Since it's one-time payoff schedule, each loan has one single due date
Paid off time	The actual time a customer pays off the loan
Past due date	How many days a loan has been past due
Age	Age of the customer
Education	Education level
Gender	Male or female

BI Model

***[i] ; Here i= Screenshot number in screenshots.pdf attached with this file.

Analysis

- The data set is a predictive analysis data set where Loan status would be the target variable [1]*. We would divide our process of mining into 4 steps.
- First being the preprocessing or data cleaning.
- Then second being the exploring the dataset for finding the input variables which are related to the target variable.
- The penultimate is building the predictive model using decision tree.
- Last being using different algorithms for a comparative study for getting the highest accuracy.

Step 1 -> Preprocessing.

- The data set had 53 values of dates which had a format different from the rest of the data in effective date column. We replaced the dates in excel just by replacing the string with dates value.
- **Stat Explorer**:-
  We use Stat Explorer node to see the chi-square and the worth of the target variable in relation to the input variables [2-3]. We obtain that the variables Effective Date, due date, Loan id and Paid-off time are not important variables in comparison to other variables [4].
- **Impute:**
  To check and find out the missing values [5]. There are 60 missing values of past due days [6]. For regression and neural networks missing values are not recommended
- **Replacement**
  Replacement enables you to remove observations from data according to specified criteria for interval variables [7].We also used replacement to enumerate categorical variables.
- **Transform variables**
  Transforming data improves model response. Transforming the data tends to stabilize variance, remove non-linearity, improve, additivity and counter non-normality. Transforming data leads to better normal fits [8].
  We have variables Age, principal and Term which are skew when seen in histograms. Now age is moderately left skew or negative skew graph [10] while principal is substantially right skew [9] or positive skew graph and Term is moderately right skewed [11]. To remove skewness of graphs we will apply Log function to substantially positive skew graph of principal and applying square root function to other two variables as they are moderately left and right skewed[13-16].

  After preprocessing we have concluded that we will have to eliminate the Effective date, Due date, Paid off time and loan id. So, we are changing their type to reject.

Step 2-> Data Exploring

- **Data partition**
  Data partition divides data into training, validation and test data set. We have selected 50% training and 30% validation and 20% test data for our dataset [20].
- **Variable clustering**
  It is used to reduce the number to something smaller by using a representative variable for a set of variables which are fairly independent or not completely independent or not highly correlated to the original data set as others are [21][22]. As we already have less numbers we are not going to use it.
- **Variable selection**
  Variable selection node is used to quickly identify the input variables that are useful for predicting the target variables [23]. It can predict using R-square or chi-Square value or both [24] [25]. For regression model we select R-square value and for classification models, we select Chi-square values where there is a categorical target variable [26].
- **Segmentation Profile**
  It can be used as a pre-screening to see which variables are related to the target variable [27] [28]. We can use segment profiling node to our data set by setting the analysis variable to segment variable and find out the related variables to target variable [29-31].

We have concluded that Principal, Age, Gender, Education, Past Due Date and Terms are important variables that could be used for predicting the target variable Loan Status.

Step 3-> Building a model

- Predict or classify the target variable, that is Loan_Status based on the input variables.

- **Decision Tree**
  An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. We used Interval Target Criterion which specifies the method of searching for and evaluating candidate splitting rules in the presence of an interval target as Variance. We used Nominal Target Criterion which our target variable as Entropy [40]. Nominal Target Criterion specifies the method of searching for and evaluating candidate splitting rules in the presence of a nominal target. We selected the maximum branch as 2 or binary and ran the decision tree.

- **Logistic Regression**
  Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes (3). We used all the three types of logistic regression [41-43]. If Backward is selected, training begins with all candidate effects in the model and removes effects until the Stay significance level or the stop criterion is met [44]. If Forward is selected, training begins with no candidate effects in the model and adds effects until the Entry significance level or the stop criterion is met [46]. If Stepwise is selected, training begins as in the Forward model but may remove effects already in the model [45]. This continues until the stay significance level or the stop criterion is met. We concluded that the forward regression is the best fit for our model.

- **Neural Networks**
  Neural Networks are a class of flexible, nonlinear regression models, discriminant models, and data reduction models that are interconnected in a nonlinear dynamic system [47]. We used Neural networks for our model with default values as Average error as selection criterion [48].

- **High Performance Bayesian network Classifier**
  It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. We used this node to classify the target variable and used Bayesian network as model type [49-50].

- **High Performance Forest**
  Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. We used Random forest for classification of our target variable [51-52].

Step 4-> Comparing different Models.

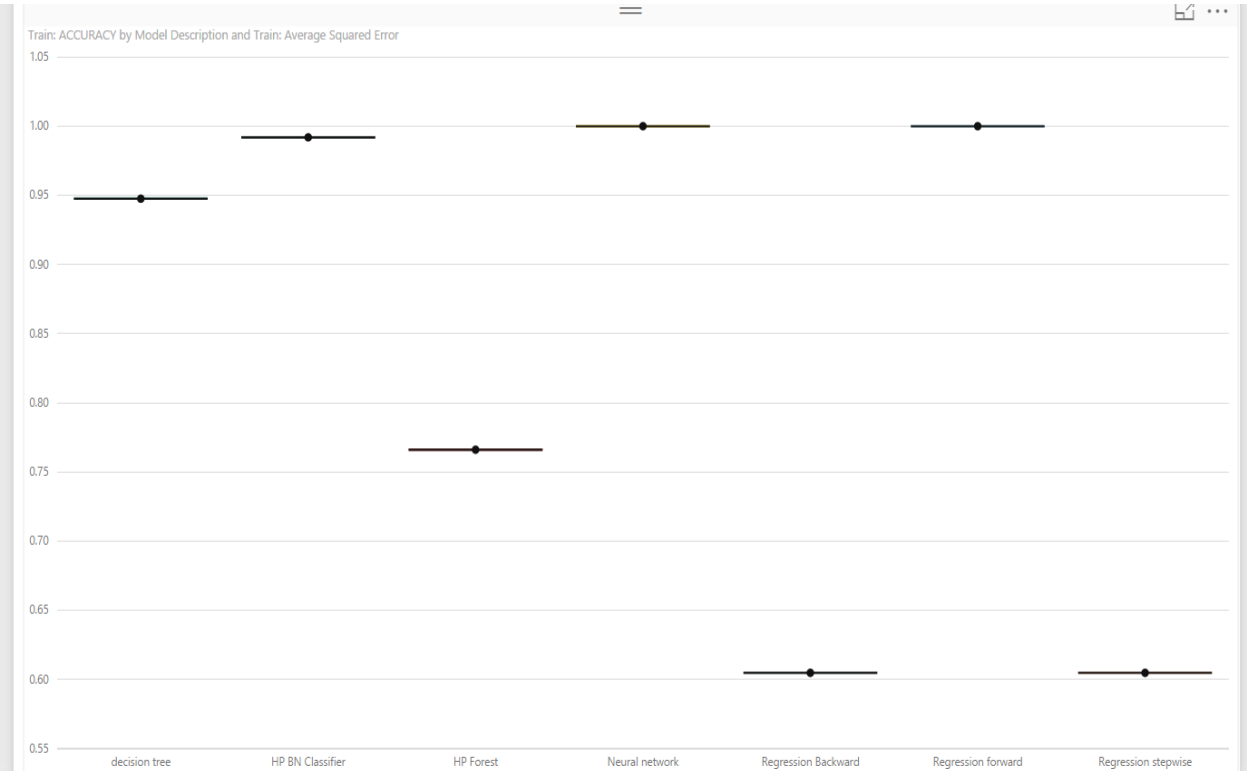
- **Model Comparison**
  Compares models and predictions from the preceding modeling nodes. We used this node to compare different classification models. The models used were Decision tree , Random Forest, Bayesian Networks, Logistic Regression(3 types) and Neural Network. The values used to compare accuracy of models are shown in table [53-55]. (Accuracy=1-missclassification rate)

Model Description	Valid: Accuracy	Valid: Roc Index	Valid: Average Squared Error
Regression forward	0.97	0.971	0.0167867961629296
HP BN Classifier	0.97	0.978	0.0718770143132775
Neural network	0.965	0.996	0.0177354881803054
decision tree	0.96	0.989	0.0185641735918744
HP Forest	0.775	0.751	0.127916396942724
Regression Backward	0.6	0.5	0.186678373222338
Regression stepwise	0.6	0.5	0.186678373222338

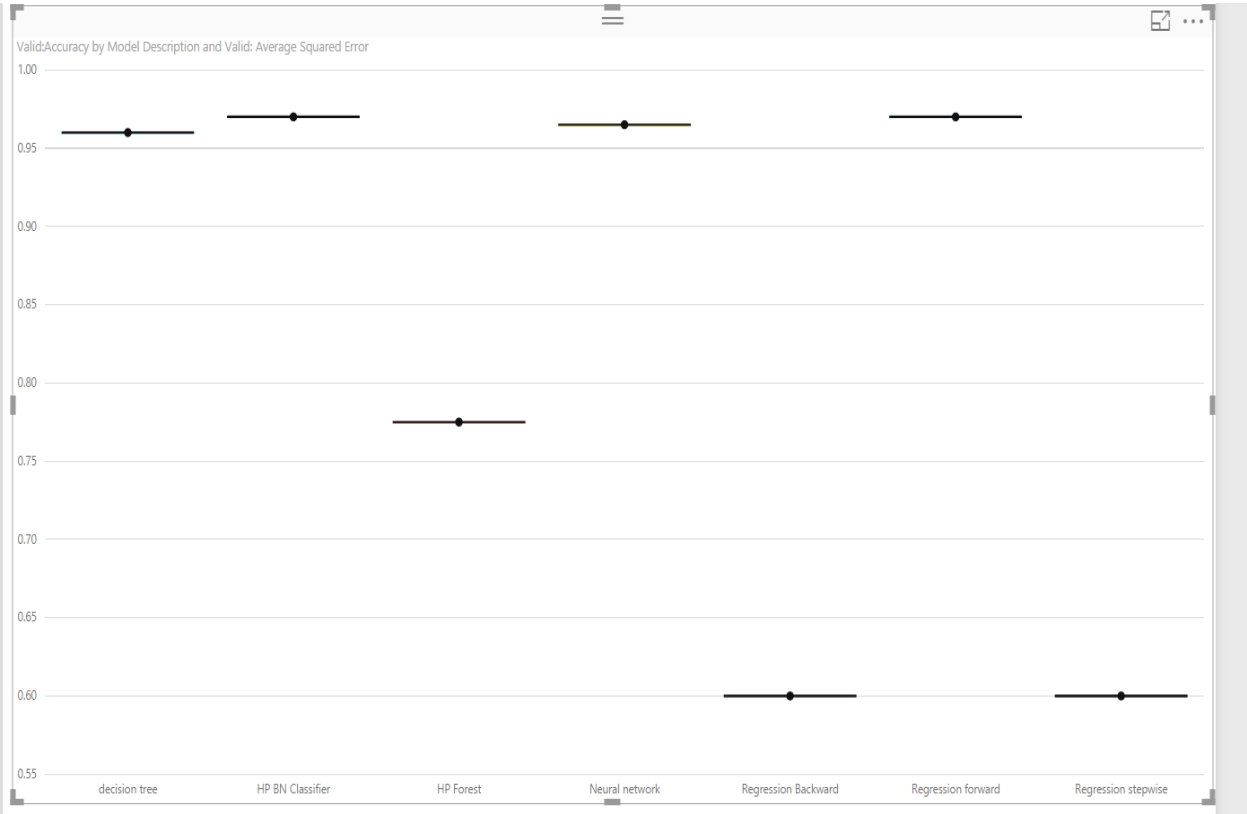
Model Description	Train: ACCURACY	Train: Roc Index	Train: Average Squared Error
Regression forward	1	1	3.49860630698337E-07
HP BN Classifier	0.991935483870968	0.999	0.0665988525631205
Neural network	1	1	0.0001258693254218
decision tree	0.94758064516129	0.983	0.0229909451046973
HP Forest	0.766129032258065	0.723	0.134097674903058
Regression Backward	0.604838709677419	0.5	0.185364637530351
Regression stepwise	0.6048387096774 19	0.5	0.185364637530351

Observations from the Dataset

We have observed from the model comparison that logistic regression (forward), Neural Network and Bayesian networks are the best fit for our data set. We can observe this from the following visualizations.



Graph : Model vs Train Accuracy.



Graph: Model vs Validation Accuracy.

Conclusions

We conclude from the observations that for our model the accuracy of logistic regression is high at training level but it reduces at validation level. The Bayesian Network accuracy is higher at the validation level so we must take this as our classification model. We must implement this model on other dataset to evaluate it in a more rigorous manner.

References

1. http://time.com/money/4701506/student-loan-defaults-record-2016/
2. https://www.kaggle.com/zhijinzhai/loandata
3. https://www.medcalc.org/manual/logistic_regression.php