

Home Equity Loan

Name: Bala Durga Sridevi Kesavarapu



**Home Equity Loan (HMEQ) Approval Analytical Model**

**Abstract:**

The goal of this assignment is to create an analytical model to determine who should be approved for a home equity loan. This model is built in such a way that the consumer credit department of a bank can automate the decision-making process for the loan approval. The model is built by using the predictive modeling tools.

The model is built in the SAS Enterprise Miner, an interface that simplifies the tasks associated with analysis. The SAS Enterprise Miner tools Sample, Explore, Modify, Model, Assess (SEMMA) will be used to create the process flow. To develop a best model, we need to perform three tasks.

1. Data Exploration: Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a dataset. It provides graphs that can be used to explore the relationships among selected variables. In this we will create the project, define a library and create an analysis diagram. After that we will analyze the data to see if any changes required like removal of data or imputing or transforming it. Any changes made to data will be noted and explained.

2. Descriptive analytics: Next step is to perform descriptive analytics to the explored data. The Descriptive analytics allows us to understand the past for generating testable hypotheses for predicting the future. A descriptive study is “concerned with and designed only to describe the existing data, without regard to causality or other hypotheses.”

3. Predictive analytics: Last step is to perform Predictive analytics to the modified dataset. Predictive analytics allows for predicting with confidence what will happen next based on what has happened so that you can make smarter decisions and improve outcomes. The goal of the predictive analytics is to use the reflect historical data, to automatically generate a model that can predict a future behavior. In this we will partition the data into three parts and then develop predictive models using the decision tree predictive method. Then the best tree will be selected by the model comparison tool.

By successfully performing these three tasks the bank will have a model which automates the process of loan approval. The detailed process will be explained in the report.

**Analytical Report:**

This report for this project is developed based on the findings of the dataset. This report will help the management to decide whether to approve a loan or not.

A dataset HMEQ is given which contains baseline and loan performance information for 5,960 recent home equity loans. The first step is to prepare the data. A library is created in SAS Enterprise Miner which has HMEQ dataset. Next, a SAS Enterprise Miner Diagram is created. We now will define the data source which is a link between an existing SAS table and SAS Enterprise Miner. To define a data source, we need to select the analysis table and define HMEQ metadata.

**Data Exploration:**

By exploring the data, we can substantially reduce the chances of erroneous results in our analysis, and we can gain insights graphically into associations between variables. In this exploration, we will look for sampling errors, unexpected or unusual data values, and interesting variable associations. We will construct interactive plots to explore the data. The dataset HMEQ can be explored by right clicking it under the data sources and select explore.

By doing this basic exploration we can find that there are some missing values in the provided data. The percentage of missing for Debt to income ratio is highest among all the variables which is 21.2% and the least 0% or no missing values for Loan and Default variables. Job category “Others” has the most loan defaulted frequency of 2338.The number of loan defaulted cases is 1189 and non-defaulted cases is 4771.

The Data Exploration can be intensively done by using the statExplore. It can be found under the Explore tab. Drag it to the workspace and connect it to the HMEQ dataset. Right click the statExplore and run it to explore the results. The following can be observed.

* The Skewness of the input variables, all inputs are right skewed, but the high level of skewness is for Derogatories variable which is 5.3% and Delinquencies which is 4.02%.
* Positively skewness or right skewed indicates that many low score and few high scores. Whereas negatively skewed or skewed left indicates that few low scores and many high scores.
* The Mean of the variable Mortgage is highest for defaulted than the other variables.
* The standard deviation of the property value is highest for defaulted than the other variables.
* The kurtosis of the property value is high for defaulted than the other variables.

Some changes need to be made to the dataset in order to filter the data and avoid the erroneous data. Changes can be made by removing, imputing, transforming or massaging data set. The focus is on missing data, range of data, max, min, distribution, skewness and outliers. Since we observed that there are missing values in the dataset, we need to replace them with reasonable values. To do this we need to drag the replacement node which is under the modify tab to the diagram workspace and connect it to the HMEQ dataset. In the left side property panel of the replacement node change the default limit method to none. Under the score property change the replacement values to missing. Now, Right click the replacement node to run and see the results.

* The replacement count of the CLNO is highest which is 305. The variable mortgage has some outliers in the data. so, we have to filter this to make it a reasonable data set.
* New columns will be added for the replaced variables. Rows with “.” Indicates the missing values.

**Descriptive analytics:**

Descriptive analytics allows us to understand the past for generating testable hypotheses for predicting future. A descriptive study is “concerned with and designed only to describe the existing data, without regard to causality or other hypotheses.” Descriptive analytics help to understand the relationship between or among data elements. The objective is to gain an understanding of what approach to take in the future. Descriptive Analytics is based on Graphs or visualization, measure of central tendency, measure of variability, measure of relative position and measure of relationship.

* Measure of central tendency: The measure of central tendency is mean, median and mode. The summary statistics of the replacement node indicates that the mortgage variable has highest mean and CLAGE variable has highest median and mode is for the variable job “other”.
* Measure of variability: The measure of variability is range, interquartile range, variance and standard deviation. By plotting a box plot graph, we can observe that the inter quartile range is 275, standard deviation is highest for the property value. With the help of box plot, it is evident that people who took loan for more than $80,000 defaulted on loan with the reason either for their loan was debt consolidation or home improvement.

**Predictive Analytics:**

Predictive analytics identifies and mathematically represents underlying relationships (not necessarily casual) in historical data in order to explain the data and make predictions, forecasts about future events. Predictive Analytics are frequently operationalized in mission critical applications and drive decisions and actions in near real time.

The next step is data imputation. For example, if an interval input contains a missing value, replace the missing value with the mean of the non-missing values for the input. This eliminates the incomplete case problem but modifies the input’s distribution. This can bias the model predictions. Making the missing value imputation process part of the modeling process allays the modified distribution concern.

Any modifications made to the training data are also made to the validation data and the remainder of the modeling population. A model trained with the modified training data is not biased if the same modifications are made to any other data set that the model might encounter (and the data has a similar pattern of missing values).

To perform imputation, the impute node which is under modify tab is dragged to the diagram workspace and then connected to the replacement node. Run the impute node and examine the results. There will be 0 missing values.

The next step is to partition the data. The imputed dataset is now portioned into 60,40% Training and validation respectively. he validation data set is used for monitoring and tuning the model to improve its generalization. The tuning process usually involves selecting among models of different types and complexities. The tuning process optimizes the selected model on the validation data. Consequently, a further holdout sample is needed for a final, unbiased assessment. The test data set has only one use: to give a final honest estimate of generalization.

Consequently, cases in the test set must be treated just as new data would be treated. They cannot be involved whatsoever in the determination of the fitted prediction model. In some applications, there may be no need for a final honest assessment of generalization. A model can be optimized for performance on the test set by tuning it on the validation set. It may be enough to know that the prediction model will likely give the best generalization possible without actually being able to say what it is. In this situation, no test set is needed.

With small or moderate data sets, data splitting is inefficient; the reduced sample size can severely degrade the fit of the model. Computer-intensive methods such as cross validation and the bootstrap have been developed so that all the data can be used for both fitting and honest assessment. However, data mining usually has the luxury of massive data sets.

After the partition of the data, the predictive model of decision tree is developed. To do this, select decision tree node under the model tab and drag it to the diagram workspace and run it with the default settings. Three interactive trees are developed by creating a split rule for each of them. Misclassification rate of each tree is observed by viewing the sub assessment tree.

|  |  |  |
| --- | --- | --- |
| **Tree name** | **Split rule** | ***Misclassification rate*** |
| Decision Tree 1 | Default | 10.6% |
| Interactive Decision Tree 1 | Split node at Delinquencies | 11% |
| Interactive Decision Tree 2 | Split node at Debt to income ratio | 11.2% |
| Interactive Decision tree 3 | Split node at Age | 11.1% |

The best tree can be selected by using the model comparison node. The model comparison node can be found under Assess tab. Drag it to the diagram workspace and connect it to the 4 decision trees we have created. Right click the model comparison node and run it to view the results.

The measure of goodness in this case is the misclassification rate. The tree which has lowest misclassification rate among the trees is selected as the best model. In this case the default tree is selected as the best model since it’s misclassification rate is lower than the other three trees.

**Model implementation:**

The best model is implemented after it is selected, this model must be put to use. The contribution of SAS Enterprise Miner to model implementation is a scoring recipe that is capable of adding predictions to any data set structured in a manner similar to the training data. We will be using internally scored dataset. Right click the Data sources in the project panel and select the Create data source. Next, add the HMEQ dataset to the library. Then change it’s role to Score.

Click the Assess tab and drag the score tool to the diagram workspace and connect it to the model comparison node. Drag the HMEQ data source into the workspace and connect it to the score node. Now run the score node and view the results. The item of greatest interest is a table of new variables added to the score data set.

**Conclusion:**

We have developed a perfect model by performing the Data exploration, Descriptive analytics and Predictive analytics. The findings of this report will help the management to decide whether to approve a loan or not. We developed a set of rules and guidelines in our model for the management to follow by providing the best analytical model. The findings are

1. people who took loan for more than $80,000 defaulted.
2. Those who defaulted, Debt-to-income ratio are lower for those who requested loan for debt consolidation than home improvement.
3. People in the job category- sales; defaulted the most if they had more than 35 trade lines.