*HUNG HUYNH-1108389*

*Southern New Hampshire University*

*CAPSTONE PROJECT*

DAT-690 (1/21/2018)

**CONTENT**

PART I: Introduction-------------------------------------------------------------------------------------------- 3

PART II: Security Goal------------------------------------------------------------------------------------------ 13

PART III: Model Creation-------------------------------------------------------------------------------------- 15

PART IV: Pilot Analysis----------------------------------------------------------------------------------------- 21

PART V: Deployment------------------------------------------------------------------------------------------- 55

PART VI: Conclusion-------------------------------------------------------------------------------------------- 56

REFERENCES------------------------------------------------------------------------------------------------------ 57

Appendix: Model deployment output--------------------------------------------------------------------- 58

**PART I: INTRODUCTION**

For this part I am going to write about how the existing data (credit/finance dataset) and the data sources could potentially provide business value for the organization. I also explain how the CRISP-DM process will enable proper execution. I will select an analytic structure (descriptive, predictive, prescriptive, or some combination) and explain how the structure fits the organization and how it could provide additional support, benefits, and values for the organization as a whole. And finally, I propose analytic tools to use in data research and modeling, e.g. what elements factor into consideration in choosing this tool. Also, I will state the problem/challenge within General Electric. I will state the purpose and type of my research or analysis and provide context for business challenge. I will also Identify stakeholder needs for data analytics, assess stakeholder needs for data analytics, apply knowledge of stakeholder needs for analytic plan and identify successes and challenges from pilot run.

General Electric (GE) is a huge, multinational corporation that consists of numerous entities and subsidiaries spread throughout the world. Based in Boston, Massachusetts and employing over 300,000 employees, it is one of the largest companies in the world. Its diverse business operations include such credit, finance. In recent years, middle management at the company has noticed a rising trend in bad credit loans from customers with high potential, combined with rising numbers defaults. As is the universal detriment of any company with regard to time and resources spent to acquire new method, GE is aware of the need to stem this tide of an inefficient nature. This paper will attempt to break down the identification of the problem into understandable segments so that the department can have a better understanding of identifying GE problem. First, consider the scenario, the financial crisis of 2008–2009. The credit branch has been asked to reassess the method used to determine if an application presents a bad credit risk. The example file, Credit\_Data, has data on 1,000 past credit applicants, described by 31 variables. The goal is to obtain a model that may be used to determine if new applicants present a good or bad credit risk by identifying factors that make an applicant at higher risk of default.

The only way to understand there is a problem in the first place is to describe and summarize the issue in relation to other time periods. Description of the issue will include data of various types of loans. One period in time will be contrasted with other time periods, in order to assess if the problem is growing, decreasing, or static. These statistics will be internal GE-based data and rely on credit data reports, surveys and compiled data. There is also a data source provided by an external data source to support the analysis it would be Credit Approval Data Set from UCI machine learning repository (<http://archive.ics.uci.edu/ml/datasets/credit+approval>). This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values. It provides a comprehensive overview of credit approval. In conjunction with the internally-based GE reports, it is possible to see if any correlations exist between customers with high potential who are in default. We can understand that there are four affect the approval decision while others have no impact. The four factors all positively affect the outcome and that as these factors increase, so does the probability that a credit card will be issued.  
The four influencing factors are:

1. Prior default,
2. Years employed,
3. Credit score, and
4. Income level.

Using the output below, we can see that the outcome values in Approved are ‘+’ or ‘-’ for whether credit had been granted or not. These character symbols aren’t meaningful as is so will need to be transformed. Turning the ‘+’ to a ‘1’ and the ‘-’ to a ‘0’ will help with classification and logistic regression models later in the analysis.

'data.frame': 689 obs. of 16 variables:

$ Male : num 1 1 0 0 0 0 1 0 0 0 ...

$ Age : chr "58.67" "24.50" "27.83" "20.17" ...

$ Debt : num 4.46 0.5 1.54 5.62 4 ...

$ Married : chr "u" "u" "u" "u" ...

$ BankCustomer : chr "g" "g" "g" "g" ...

$ EducationLevel: chr "q" "q" "w" "w" ...

$ Ethnicity : chr "h" "h" "v" "v" ...

$ YearsEmployed : num 3.04 1.5 3.75 1.71 2.5 ...

$ PriorDefault : num 1 1 1 1 1 1 1 1 1 0 ...

$ Employed : num 1 0 1 0 0 0 0 0 0 0 ...

$ CreditScore : num 6 0 5 0 0 0 0 0 0 0 ...

$ DriversLicense: chr "f" "f" "t" "f" ...

$ Citizen : chr "g" "g" "g" "s" ...

$ ZipCode : chr "00043" "00280" "00100" "00120" ...

$ Income : num 560 824 3 0 0 ...

$ Approved : chr "+" "+" "+" "+" ...

To start with, we will use the summary() function to see the descriptive statistics of the numeric values such as min, max, mean, and median. The range is the difference between the minimum and maximum values and can be calculated from the summary() output. For the B variable, the range is 66.5 and the standard deviation is 11.9667.

Age Debt YearsEmployed CreditScore Income

Min. :13.75 Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0

1st Qu.:22.58 1st Qu.: 1.000 1st Qu.: 0.165 1st Qu.: 0.000 1st Qu.: 0

Median :28.42 Median : 2.750 Median : 1.000 Median : 0.000 Median : 5

Mean :31.57 Mean : 4.766 Mean : 2.225 Mean : 2.402 Mean : 1019

3rd Qu.:38.25 3rd Qu.: 7.250 3rd Qu.: 2.625 3rd Qu.: 3.000 3rd Qu.: 396

Max. :80.25 Max. :28.000 Max. :28.500 Max. :67.000 Max. :100000

NA's :12

[1] 11.9667

The table below shows the correlation between all of the variables. The diagonal correlation values equal 1.000 because each variable is perfectly correlated with itself. To read the table, we will look at the data by rows. The largest value in the first row is 0.396 meaning age is most closely correlated with YearsEmployed. Similarly, Debt is mostly correlated with YearsEmployed.

Age Debt YearsEmployed CreditScore Income

Age 1.000 0.202 0.396 0.186 0.019

Debt 0.202 1.000 0.301 0.271 0.122

YearsEmployed 0.396 0.301 1.000 0.327 0.053

CreditScore 0.186 0.271 0.327 1.000 0.063

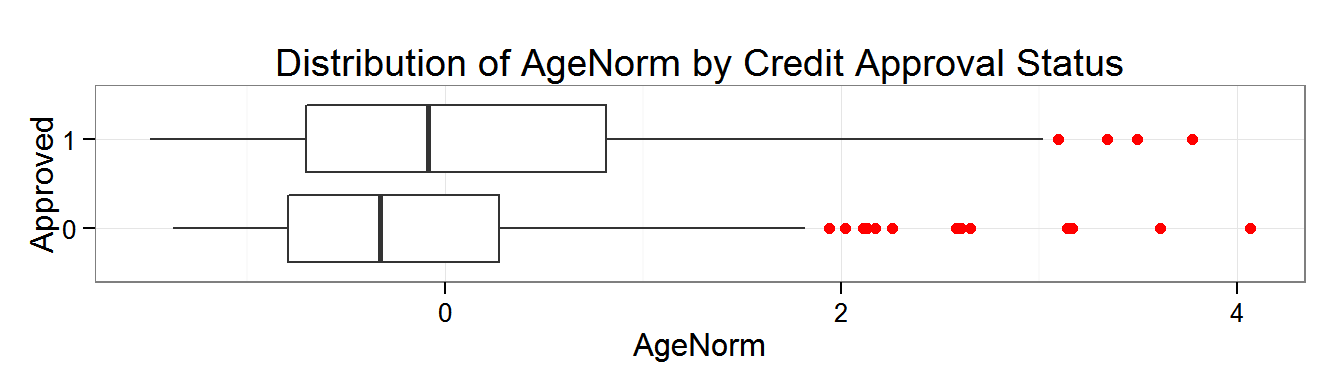
Income 0.019 0.122 0.053 0.063 1.000

We can use this information to create a linear regression model between the two variables. The model produces the two coefficients below: Intercept and YearsEmployed. These coefficients are used to predict future values. The YearsEmployed coefficients is multiplied by the value for YearsEmployed and the intercept is added.

(Intercept) YearsEmployed

28.446953 1.412399

We’ll use a boxplot showing the mean value for each group and the quartiles. We can tell from the boxplot, that the median of the two groups is slightly different with the age of approved applications being slightly closer to the mean than the denied applications. We can also see that the interquartile range is greater on the ‘Approved’ than the others. We can interpret these facts as the credit applicants with lower Age values are less likely to be granted credit, however there are several outlying applicants with high values that still were not granted credit.



**Develop Research Questions**

1. Is there a correlation between Age, Income, Credit Score, and Debt levels and the credit approval status? Can this relationship be used to predict if a person is granted credit? If yes, does the relationship indicate reasonable risk management strategies?
2. Ethnicity is a protected status and the decision to approve or deny an application cannot be based on the applicant’s ethnicity.  
   Is there a statistically significant difference in how credit is granted between ethnicities that could indicate bias or discrimination? Contrarily, could the difference indicate a business opportunity?

There were 3 significant numeric variables- YearsEmployedLog, CreditScoreLog, and IncomeLog. Remember that these 3 variables are the logarithmic transformations of YearsEmployed, CreditScore, and Income. The other numeric variables fed into the model did not have a significant impact on the approval decision. This means that Age and Debt did not have an influence on the final credit approval outcome. the coefficients for the 3 variables, we can see that they are all positive. This means that the probability of getting approved for a credit card increases as the values for YearsEmployedLog, CreditScoreLog and IncomeLog increase. These relationships make sense for a credit application so there’s no exception taken. While we’d expect Ethnicity does not have an impact on the approval decision, we can do a simple Chi-Squared test to gain additional confidence for compliance testing. Other variables such as age, sex, or ethnicity did not have an influence on whether the application was denied. A Chi Squared test for independence validated our conclusion Ethnicity and Approval status are independent.

Pearson's Chi-squared test

data: tbl[2:3]

X-squared = 18.5161, df = 8, p-value = 0.01767

(Analysis of Credit Approval Data, Ryan Kuhn)

A method of classifying these high-potential bad loans into those that will probably in default, versus those who probably will not, will aid the organization in better retaining business. Decision tree learning can offer a concrete visual to whoever analyzes the trends of growing bad loans. The costs, both financial and temporal, can be decreased when a good and comprehensive prospectus exists for “red flag” loans. A valid prediction technique can further aid in having an empirical evidence-based analysis of which company is losing critical money due to bad loans. With solid predictive and classification techniques available, GE departments will be able to prevent, if not most, further loan defaults. Regression analysis offers a solid method of making a prediction of which staff members will be most prone to changing direction toward the “positive”.

We’re going to be starting another data extraction project soon for a different system and a different sponsor. Here’s what we undertake that journey, and if stakeholders sponsoring initiatives, it may be useful to:

* Confirmation bias is real. Our stakeholders are talented people who have risen high up the management chain over their careers. And, like ours, it’s likely that some of stakeholders did so for decades without any data analysis supporting their management decisions. They’re used to trusting their gut and may expect the data to exclusively support their ideas.
* Praise in public, criticize in private. If our sponsor (or anyone on the project) has a hypothesis that doesn’t prove out, the public project meeting is not the place to unveil it. Meet with them ahead of time to fill them in and get their buy-in on a follow-up inquiry. This gives them buy-in to these efforts and lets them look clever, not wrong, in the public meeting.
* Teach the process of analysis, not just the outcomes. We focused too much on how exciting our potential “wins” could be with our data. Our sponsors only had the information they’d gotten from colleagues and conferences which was “executive-level” summary information. A broader understanding of the process could lead to more inquisitive stakeholders who will be more supportive of organic inquiries.
* “Do something” with the data. We won’t just wait for our stakeholders to tell us what they want to know. As we bring new data sets online we’ll proactively mine them for interesting (and, perhaps, flattering) insights. This will give us some political credit to draw from if we need to defend an unpopular insight. It also means that our stakeholders will be introduced to the data on terms favorable to us, instead of a roll of the dice on the merits of their own ideas

In ﬁghting the effects of the ﬁnancial crisis of 2007 the country enacted the Credit Institutions Financial Support Scheme 2008. In section 32 the scheme states:

“In order to promote the public interest, a covered institution shall, at the direction of the Minister, take all reasonable steps to appoint at least one but no more than two non-executive directors to its board from a panel approved by the Minister during the period of the guarantee. The covered institution shall remunerate those non-executive directors. The Minister will also have the right to appoint persons to observe all meetings of the remuneration, audit, credit and risk committees of a covered institution. Such observers shall have the right to attend all meetings and have access to necessary committee papers and other relevant information”.

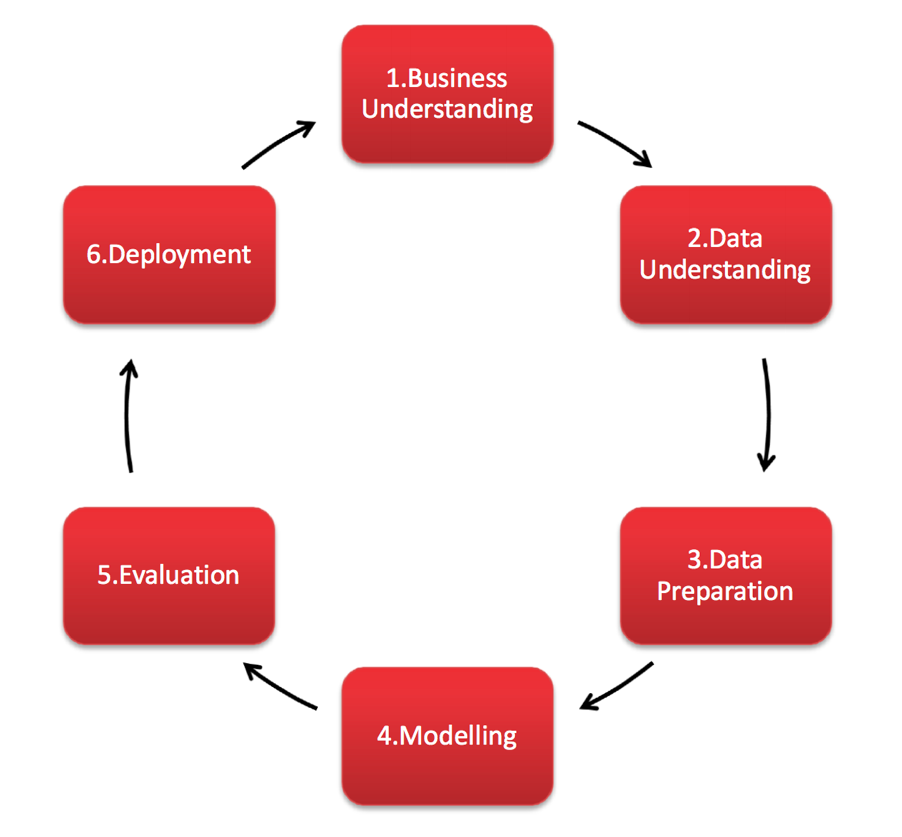
Public interest directors are one way of stakeholder governance allowing external stakeholders access to corporate decision making in order to prevent future crises from happening. This paper looks beyond legal mechanisms and investigates voluntary mechanisms of stakeholder governance applied by companies. The critical question we address by looking at voluntary stakeholder governance mechanisms is: how do stakeholders inﬂuence corporate decision making and help to align the worldviews of those inside and outside the organization?

The SME (Subject Matter Expert) has the following responsibilities in credit/finance management: (par thi)

* Support the definition of business processes
* Determine and support the implementation of a business policy, generally by providing the following:
  + the contents for the business rules that enforce the policy;
  + the process contexts in which the rules are applied.
* Oversee the execution of that policy via business rules applied. Such oversight includes confirming that the implemented rules fully and faithfully correspond to the intended policy.
* Once Rule Writers have created the first set of rules, the SME reviews the rules, and the rule flow to give feedbacks on the logic and pattern used.
* Review the results of testing and simulation
* Manage business vocabulary
* Resolve business issues relating to business rule execution.
* Be accountable for the quality of the business rule
* Approve major changes to business rule

In term of skill and competencies, the SME has a strong business knowledge and experience, some management skill, effective communication, leadership, decision making skills.

The CRISP-DM (Cross-Industry Standard Process for Data Mining) method will be utilized in this project. The process involves breaking down the analysis into Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The aspect of business understanding is about credit application. We need to know what caused a financial crisis. Understanding this entails describing the data and confirming (or not) that there is indeed an issue and suspicions are confirmed. Data understanding and preparation like before we will examine the dataset for how it will carry out based on the variables from analyses. This entails comprehending the internal credit staffing reports and making sure there is compatibility with a descriptive and predictive analytic tool. Modeling involves choosing the correct approach to manipulating the data to produce meaningful patterns. Modeling goes through a database design (conceptualize, logical and physical). Then we evaluate the model. If all go well we will be able to deploy it to present the management department and employees.



CRISP-DM

I will use descriptive (see details of generated analysis of data), decision tree (tree split on input variables) and regression (for correlation and trend for ROI) structures for the project. Also I will use PCA, ROC for model evaluation how model performs, and a little bit hypothesis testing for my model variables. With these structures, I can:

* Documenting the types and structure of the business data (logical modeling),
* Analyzing and mining business data to identify patterns and correlations among the various data points,
* Mapping and tracing data from system to system in order to solve a given business or system problem,
* Design and create data reports and reporting tools to help business executives in their decision making,
* Perform statistical analysis of business data.

**PART II: SECURITY GOAL**

In this part, the following details are covers:

1. Evaluate the data source security requirements for the organization and current architecture. Where is the organization lacking in terms of security requirements and ethical strategies? What are some of the security and privacy strengths and weaknesses of the current architecture?
2. Recommend a strategy for securing data. Given the weaknesses that you have identified, how would you improve the security and privacy strategies used? How would your new structure address these issues in an ethical manner? What software applications may help to improve the security standards?

The credit company cannot manage its customer data like how many customers it has and the quality of data is poor. There is no single source of agent data for the entire company, and much data about agents and customers is repeated in the various source systems. So, when it comes to business intelligence it cannot generate a clear solution for decision making. Furthermore, the technology (hardware, software, and network) is not up to day and it causes slow workflow. To better security and privacy of data system it need to adapt CRISP-DM, data security and privacy.

The company data warehouse failed because it created inefficient, vulnerable data in a timeliness. And for it administrative process, it’s costly and inefficient because many mid-level managers in business units have developed separate small groups of technical staff since the IT department is not considered to be extremely competent. It lacks of data security. There is no common security of data and no management of business rules or other forms of metadata management. Data security team should choose one of the strategy and follow closely in practice. This will make their objectives and activities better understand. They should understand their role in business functions to add true measurable and accepted value to respond to data threat. They must build a completed and secured data system and adapt new tools. This will create better customers data and so reliable.

We answer what ethical issues may arise during implementation of this strategy? The credit company should have a code of ethics. Since it is a financial institution it has its own code, a set of general guidelines to encourage employees to behave ethically and responsibly. For example, most independent accountants are members of the American Institute of Certified Public Accountants (AICPA) and must abide by the AICPA Code of Professional Conduct. Accountants who are members of the Institute of Management Accountants are bound by the Standards of Ethical Conduct of Management Accountants. A company code of ethics is useful only when the company’s actions are consistent with it. Only then can it be followed consistently within the company. Once a code of ethics consistent with the actions of the company is established, it needs to be communicated across the company. The code of ethics should be reviewed on a periodic basis to ensure that it deals with current issues facing the employees.

Finally, the tools we can use are from EMC and Baracuda. The benefits are:

* **​**Improve the effectiveness of your security operations center with industry-leading and independently top-rated products and services for advanced threat detection and cyber incident response.
* ​Expertise and innovation in applying industry best practices and frameworks from NIST, SANS, US-CERT and VERIS.
* ​Alleviate burdens on security operations teams by giving them tools, technologies and capabilities built to ease and prioritize their work.
* ​Battle-tested professional services team with the most real-world experience dealing with advanced cyber threats in the most rigorous business environments.
* ​Best practices honed over the course of 35 years and thousands of engagements, including in 370 Fortune 500 companies and 48 of the Fortune 50.

**PART III: MODEL CREATION**

In this part I am going to address Model Creation Applicability and Model Creation Value. First, I would like to evaluate existing data analytic strategies in terms of their use for data model creation within the chosen organizational environment.

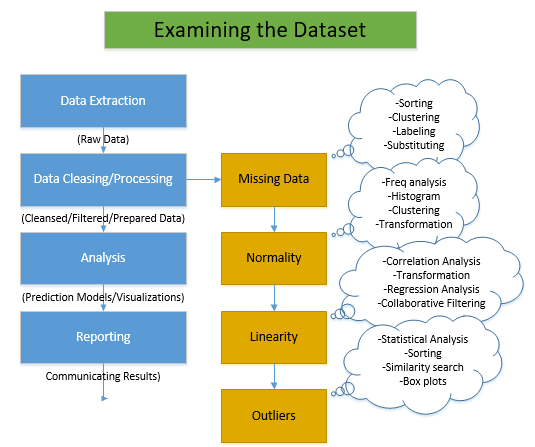
Then the values of the strategies that there’s more data available to finance departments now than ever before. The business world’s shift to digital systems has opened the door for firms to track data on everything, from industry patterns, to credit and payment terms, sales, revenue, costs, and more. This is both an incredible opportunity and new challenge for finance professionals. The company uses predictive analytics to manage its business. To do so, they need data on how much to pay them, how to benchmark the value they provide and manage any risk involved. This is especially important as companies become more global and may run into issues with managing terms of pay across multiple international borders. As the complexities for organizations grow, the need for data and information from the finance team becomes even more vital for success.

Finance and credit leaders can get started with predictive analytics by first using the data that they already have access to. Think of all the reports, income statements, balance sheets and statement of cash flows that you have—those are still important and won’t go anywhere.

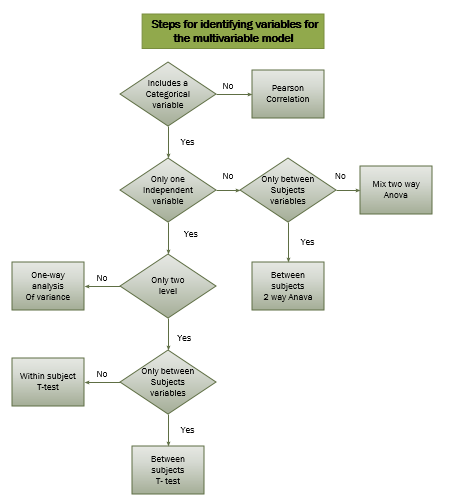
**Flowcharts:**

The following diagrams are for the project development. The model creation can adapt the charts to follow.

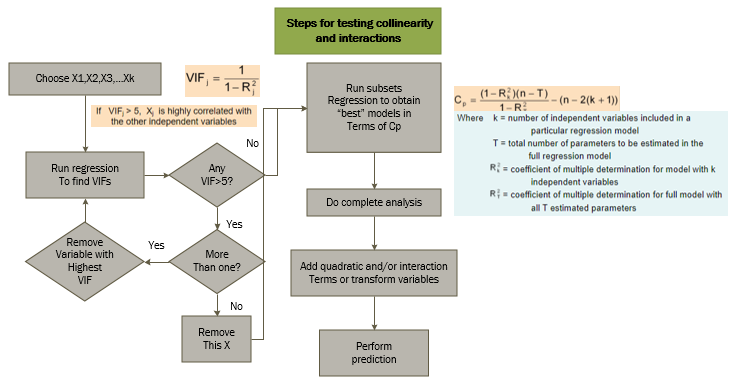
* To prepare data first we examine the data then clean it up, e.g. missing values, outliers… We can sort data, label it and substitute missing values something meaningful. We also check for normality by analyze frequency, create histogram, clustering and transform data. Then we check for linearity with correlation, regression and filtering data. Finally use statistical analysis and boxplot to remove outliers.



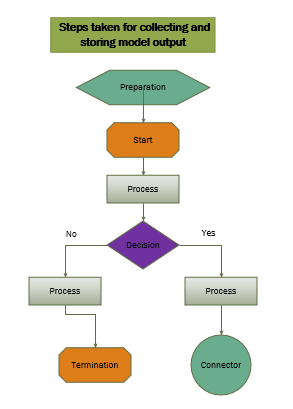
* In the same we check for multivariable. If a categorical variable check for one independent variable. If it’s two level then check for between subject variables and use T-test between subject variables, if not use T-test within subject variable. If no categorical variable then uses Pearson correlation. If no independent variables check for only between subject variables and two way Anova. If it’s one level check for one way analysis of variance.



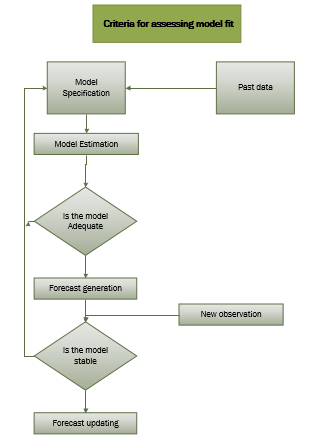
* To reduce number of variables, check for collinearity using VIFs. If VIF values for a variable is greater than 5 then remove that variable, otherwise keep it.



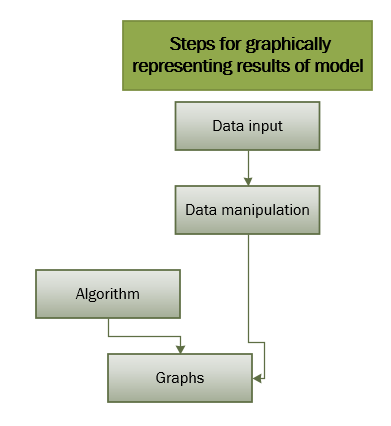
* To collect and store model output we can first prepare the model then start the process. During this phase, we can make decision. If it’s a successful decision we can make connector to other process. Otherwise terminate the program.



* To access model fit, we first specify and estimate the model. If the model is adequate then generate a forecast. If the model is stable we can update forecast. If no we have to go back to specification.



* For generating graphical, first we input data and do some manipulation with data. Then we create an algorithm to create graphs.



**PART IV: PILOT ANALYSIS**

In this part I will identify data sources and analytic structures that generate business value, e.g. additional data sources, both internal and external, could support the organization and add value in conjunction with the selected data structure. Then I also create a pilot plan in which the strategy will be implemented and tested using the available data, e.g. I will run the available data through architecture to ensure that the process is smooth and the results are as needed.

For pilot plan, I include more detailed explanation of these stages: Data preparation, Hypothesis and modeling, Evaluation and Interpretation, Deployment. This will ensure the pilot plan is transferable to related alternative proposals. (CRISP-DM process). I will use R to analyze the dataset. With R, I use it to show the trends. R programming is good for business. All of R’s programming libraries are free, but Revolution Analytics makes its business from its service packages, which give customers access to the libraries the company develops in-house. These commercial libraries are suitable for corporate customers who frequently deal with large amounts of data, in the terabyte range. With R, we can create cool graphics. “I think the number one value to businesses [in using R] is access to talent,” says Smith. “So many businesses now are doing much more with data, especially with the big data revolution and doing much more with analytics. And because they’re hiring people coming out of school. They know R already.” (Smith 5/2014). I will use decision tree (tree split on input variables) and regression (for correlation and trend for ROI) structures for the project. Also, I will use PCA, ROC for model evaluation how model performs, and a little bit hypothesis testing for my model variables. With these structures, I can:

* Documenting the types and structure of the business data (logical modeling),
* Analyzing and mining business data to identify patterns and correlations among the various data points,
* Mapping and tracing data from system to system in order to solve a given business or system problem,
* Design and create data reports and reporting tools to help business executives in their decision making,
* Perform statistical analysis of business data.

For data preparation, the dataset contains only one table without any missing value. The data is very clean and there is no outlier. This analysis demonstrates several analytic techniques to examine one company’s decision to approve or deny credit card applications. The final model created out of this analysis is a combination of a logarithmic regression model and classification and regression tree (CART) model. This model was able to predict the outcome of a credit applications with 84% accuracy which was significantly better performance than the baseline model. Classification and Regression Trees (CART) can be used for similar purposes as logistic regression. They both can be used to classify items in a dataset to a binary class attribute. The trees work by splitting the dataset at series of nodes that eventually segregates the data into the target variable. The models are sometimes referred to as decision trees because at each node the model determines which path the item should take. They have an advantage over logarithmic regression models in that the splits or decision are more easily interpreted than a collection of numerical coefficients and logarithmic scores. The goal is to obtain a model that may be used to determine if new applicants present a good or bad credit risk by identifying factors that make an applicant at higher risk of default. With this dataset, we will examine some relevant variables to determine of the research questions. For example, we can use credit history and the purpose of credit (what purpose customers apply credit for the most). We can also find the age range, and marital status to find what group apply for credit the most. “DEFAULT”. “MALE\_DIV”, “OWN\_RES”, “HISTORY”, and “DEFAULT” are all terms of the dataset that define relationships between entities and therefore factors that are involved in credit risk or default. Also, we would ask what sort of properties applicants have when apply for credits… The goal is to use as many as of these variables like number of checking accounts, new and old car indicators, employment, job nature, etc. to build a good predictive analytics model that can be used to find the credit worthiness. For example, how lenders give customers credit loans: When you apply for a loan or other type of credit, such as a credit card, the lender has to decide whether or not to lend to you. Creditors use different things to help them decide whether or not you are a good risk.

* how your credit rating is decided
* what information a creditor can find out about you to help them decide whether to lend to you
* what you can do if you are refused credit, including how to correct wrong information on your credit reference file
* how to get a copy of your credit reference file
* how fraud can affect your credit rating
* how to get credit if you’ve got a low credit score.

**ANALYSIS WITH R**

**Clean Data:**

credit\_data <- read.csv("~/Documents/SNHU/DAT-650-Q1051 Advanced Data Analytics/Credit\_Data.csv")

is.na(credit\_data) #remove missing values

sum(is.na(credit\_data))

mean(is.na(credit\_data))

*> sum(is.na(credit\_data))*

*[1] 0*

*> mean(is.na(credit\_data))*

*[1] 0*

So, the data is clean because sum of missing values is 0 and mean is 0. There are no missing values.

**VIF to Reduce Variables:**

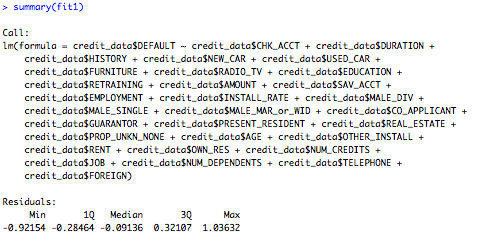
install.packages("car")

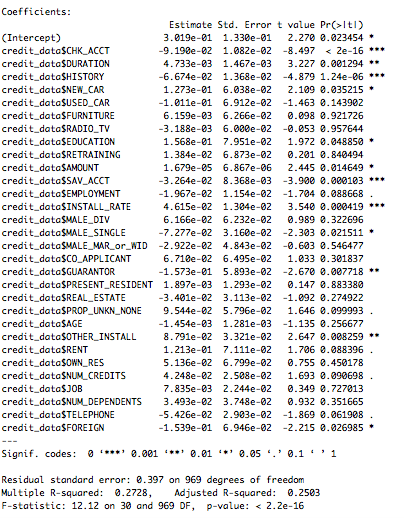
library(car)

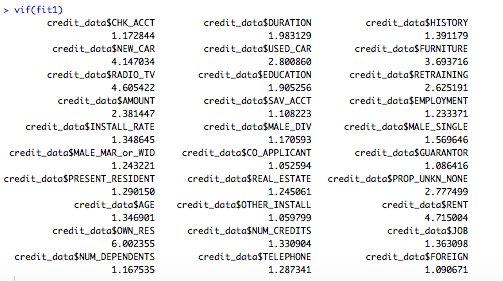
fit1<- lm(credit\_data$DEFAULT~credit\_data$CHK\_ACCT + credit\_data$DURATION + credit\_data$HISTORY + credit\_data$NEW\_CAR + credit\_data$USED\_CAR + credit\_data$FURNITURE + credit\_data$RADIO\_TV + credit\_data$EDUCATION + credit\_data$RETRAINING + credit\_data$AMOUNT + credit\_data$SAV\_ACCT + credit\_data$EMPLOYMENT + credit\_data$INSTALL\_RATE + credit\_data$MALE\_DIV + credit\_data$MALE\_SINGLE + credit\_data$MALE\_MAR\_or\_WID + credit\_data$CO\_APPLICANT + credit\_data$GUARANTOR + credit\_data$PRESENT\_RESIDENT + credit\_data$REAL\_ESTATE + credit\_data$PROP\_UNKN\_NONE + credit\_data$AGE + credit\_data$OTHER\_INSTALL + credit\_data$RENT + credit\_data$OWN\_RES + credit\_data$NUM\_CREDITS + credit\_data$JOB + credit\_data$NUM\_DEPENDENTS + credit\_data$TELEPHONE + credit\_data$FOREIGN)

summary(fit1)

vif(fit1)

**

**



All 31 variables are free from multicollinearity. All calculated VIFs are less than 5.

**Decision Tree:**

The pilot testing for my decision tree generated from R is below:

credit\_data <- read.csv("~/Documents/SNHU/DAT-650-Q1051 Advanced Data Analytics/Credit\_Data.csv")

# Regression Tree Example

library(rpart)

# grow tree

fit <- rpart(DEFAULT~CHK\_ACCT+ DURATION +HISTORY+ NEW\_CAR +USED\_CAR + FURNITURE+ RADIO\_TV+ EDUCATION+ RETRAINING+ AMOUNT+ SAV\_ACCT+ EMPLOYMENT+INSTALL\_RATE + MALE\_DIV+MALE\_SINGLE+MALE\_MAR\_or\_WID+CO\_APPLICANT+GUARANTOR+ PRESENT\_RESIDENT + REAL\_ESTATE+ PROP\_UNKN\_NONE+ AGE +OTHER\_INSTALL+RENT+ OWN\_RES+ NUM\_CREDITS+ JOB+ NUM\_DEPENDENTS +TELEPHONE+ FOREIGN,

method="anova", data=credit\_data)

printcp(fit) # display the results

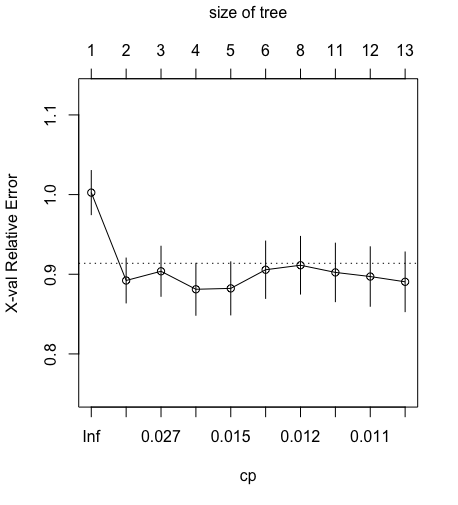
plotcp(fit) # visualize cross-validation results

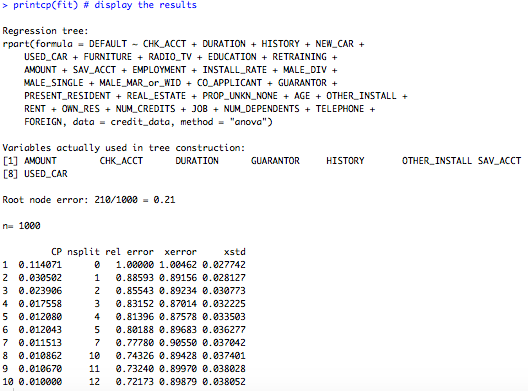
summary(fit) # detailed summary of splits

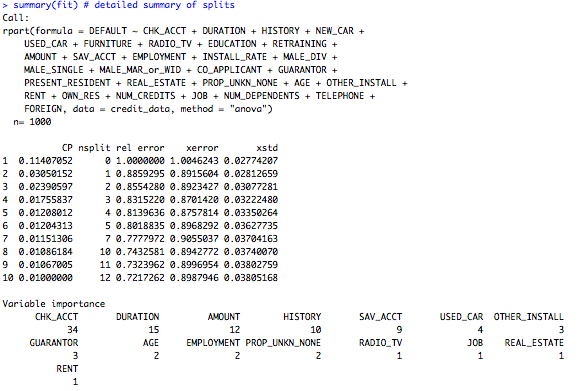
# plot tree

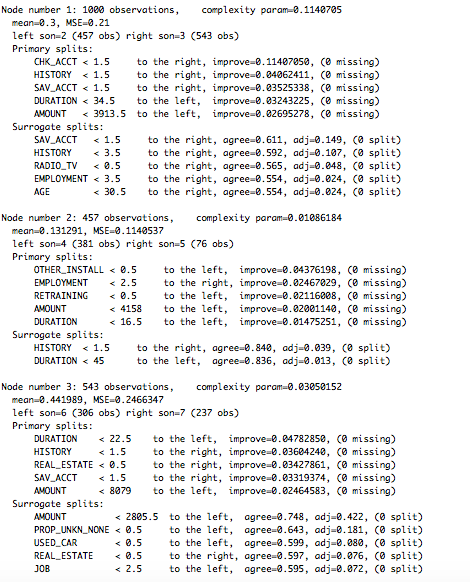
plot(fit, uniform=TRUE, margin= 0, main="Classification Tree for credit\_data")

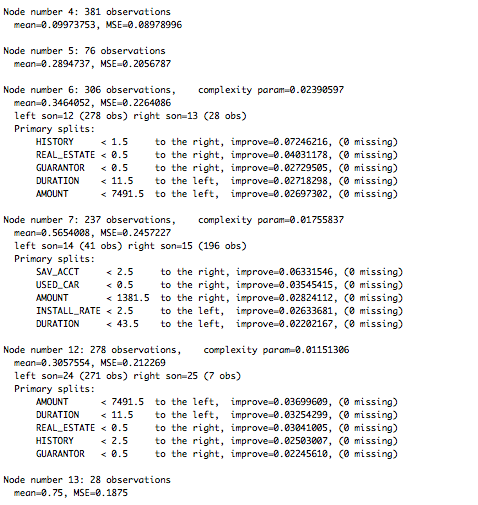
text(fit, use.n=TRUE, all=TRUE, cex=.5)

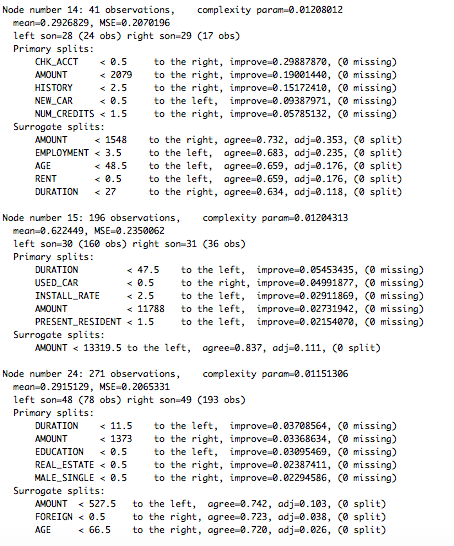


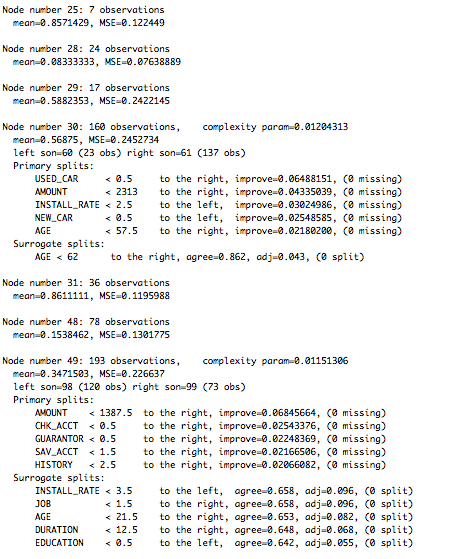


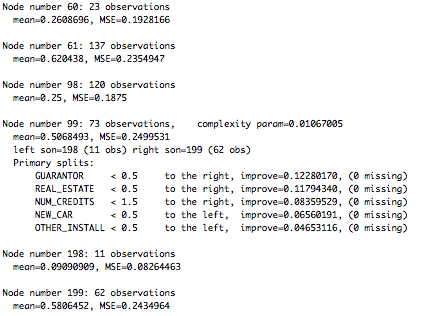


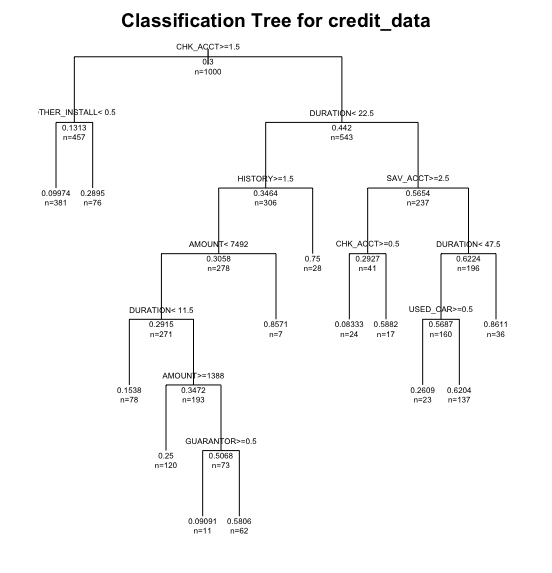












**Boxplot:**

Now, let see a boxplot for loan types. The R code is below:

#Boxplot for loan types

df<-credit\_data[, c("AMOUNT", "NEW\_CAR", "USED\_CAR", "FURNITURE", "RADIO\_TV", "TELEPHONE", "EDUCATION", "RETRAINING")]

X<-data.frame(df)

head(df)

install.packages("reshape2")

require(reshape2)

df.m <- melt(X, measure.vars = 2:8)

head(X)

head(df.m)

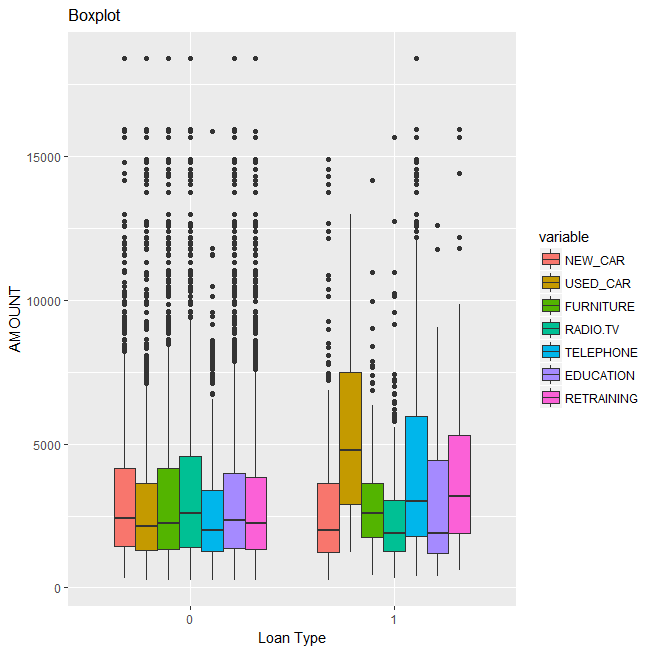
str(df.m)

View(df.m)

install.packages("ggplot2")

library(ggplot2)

ggplot(data=df.m, aes(x = factor(value), y = AMOUNT, fill=variable)) + geom\_boxplot() + ggtitle("Boxplot") + xlab("Loan Type") + ylab("AMOUNT")



Loan types USED\_CAR and TELEPHONE have highest default.

**The Confusion Matrix:**

#Logistic regression Model

library(nnet)

mymodel<-multinom(credit\_data$DEFAULT~., data=credit\_data)

#Misclassification Rate

p<-predict(mymodel, credit\_data)

tab<-table(p, credit\_data$DEFAULT)

tab

*p 0 1*

*0 628 136*

*1 72 164*

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives insight not only into the errors being made by your classifier but more importantly the types of errors that are being made.

This gives us:

* “**true positive**” for correctly predicted event values. (628)
* “**false positive**” for incorrectly predicted event values. (136)
* “**true negative**” for correctly predicted no-event values. (164)
* “**false negative**” for incorrectly predicted no-event values. (72)

We can summarize this in the confusion matrix as follows:

|  |  |
| --- | --- |
| 1  2  3 | event no-event  event true positive false positive  no-event false negative true negative |

**ROC and AUC:**

#Reciever Operating Characteristic (ROC) Curve

install.packages("ROCR")

library(ROCR)

pred<-predict(mymodel, credit\_data, type='prob')

pred<-prediction(pred, credit\_data$DEFAULT)

roc<-performance(pred, "tpr", "fpr")

plot(roc, colorize=T, main="ROC Curve")

abline(a=0, b=1)

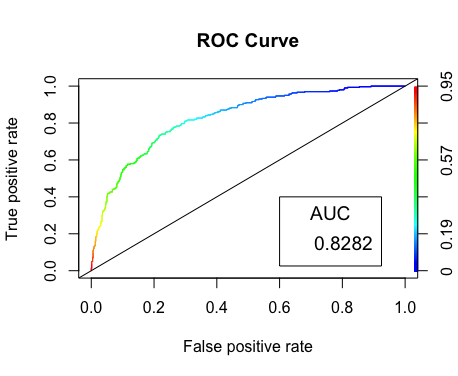
#Area Under the Curve

auc<-performance(pred, "auc")

auc<-unlist(slot(auc, "y.values"))

auc<-round(auc, 4)

legend(.6, .4, auc, title="AUC", cex=1.2)



The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target=1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target=1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases. The diagonal line is for a random model.

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1. My AUC=83% so my model performs good.

**PCA:**

Now we are using R to run PCA (Principal Component Analysis). PCA can reduce the dimensionality of the dataset to help identify those variables, which explain the most variance in the observations. These variables are the most influential in the dataset, and should be considered in the selection of KRIs.

# After rearrange data to group variables and saved as Credit\_Data2.csv

credit\_data <- read.csv("~/Documents/SNHU/DAT-650-Q1051 Advanced Data Analytics/Credit\_Data2.csv")

head(credit\_data) #show sample data

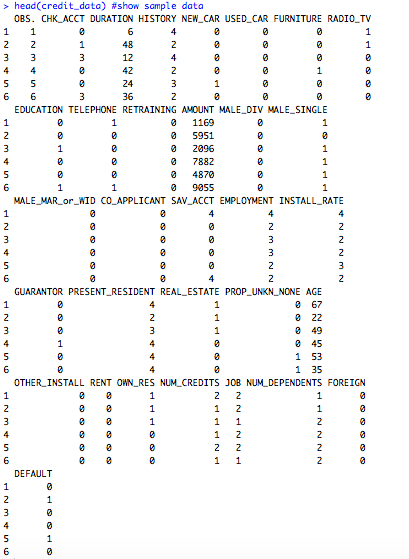
dim(credit\_data) #check dimensions

str(credit\_data) #show structure of the data

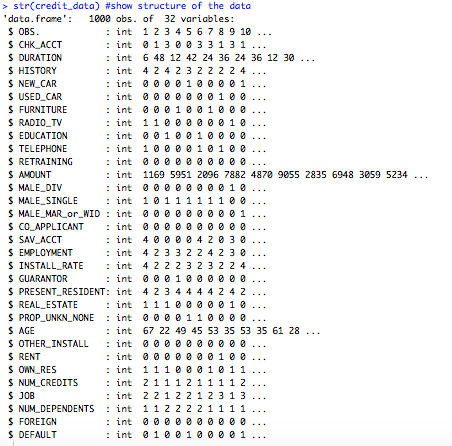
sum(credit\_data)

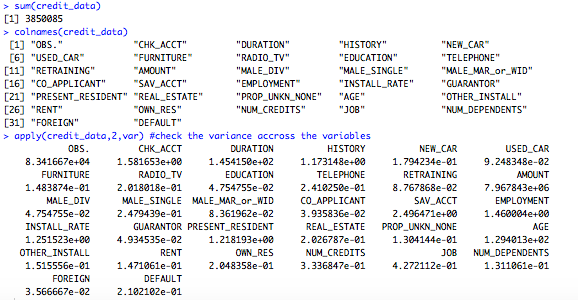
colnames(credit\_data)

apply(credit\_data,2,var) #check the variance accross the variables



**

**

**

For Loan Types: NEW\_CAR, USED\_CAR, FURNITURE, RADIO\_TV, EDUCATION, TELEPHONE, RETRAINING

pca =prcomp(credit\_data[1:1000, 5:11]) #applying principal component analysis on credit\_data

par(mar = rep(2, 4)) #plot to show variable importance

plot(pca)

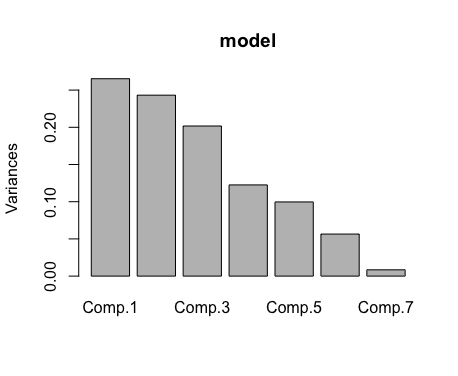
'below code changes the directions of the biplot, if we donot include

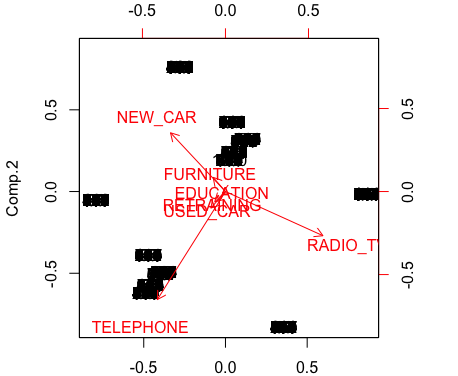
the below two lines the plot will be mirror image to the below one.'

pca$rotation=-pca$rotation

pca$x=-pca$x

biplot (pca , scale =0) #plot pca components using biplot in r



**

NEW\_CAR, RADIO/TV and TELEPHONE are used for predicting DEFAULT

For types of loaners and properties: MALE\_DIV, MALE\_SINGLE, MALE\_MAR\_or\_WID, CO\_APPLICANT, SAV\_ACCT, EMPLOYMENT, INSTALL\_RATE, GUARANTOR, PRESENT\_RESIDENT, REAL\_ESTATE PROP\_UNKN\_NONE

pca =prcomp(credit\_data[1:1000, 12:23]) #applying principal component analysis on credit\_data

par(mar = rep(2, 4)) #plot to show variable importance

plot(pca)

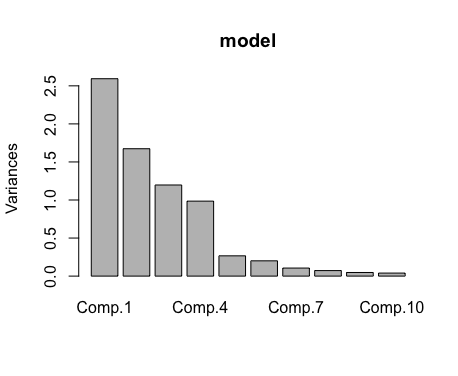
'below code changes the directions of the biplot, if we donot include

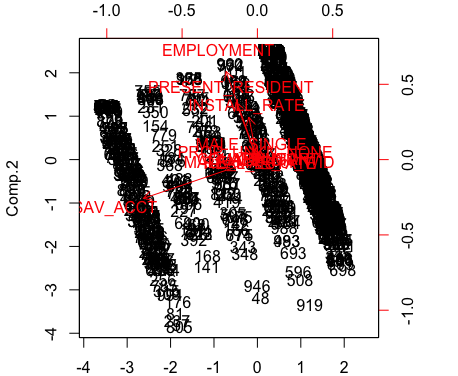
the below two lines the plot will be mirror image to the below one.'

pca$rotation=-pca$rotation

pca$x=-pca$x

biplot (pca , scale =0) #plot pca components using biplot in r





EMPLOYMENT, SAV\_ACCT & PRESENT\_RESIDENT are used for predicting DEFAULT

For others: OTHER\_INSTALL, RENT, OWN\_RES, NUM\_CREDITS, JOB, NUM\_DEPENDENTS, FOREIGN

pca =prcomp(credit\_data[1:1000, 25:31]) #applying principal component analysis on credit\_data

par(mar = rep(2, 4)) #plot to show variable importance

plot(pca)

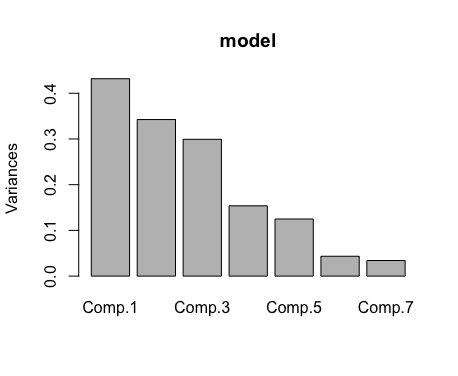
'below code changes the directions of the biplot, if we donot include

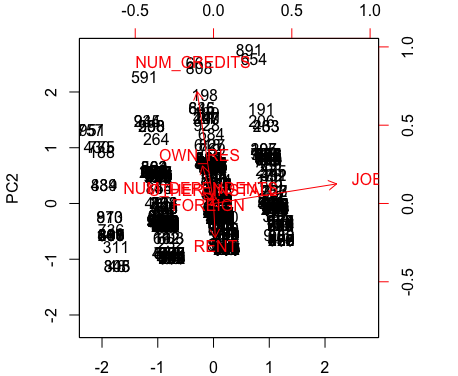
the below two lines the plot will be mirror image to the below one.'

pca$rotation=-pca$rotation

pca$x=-pca$x

biplot (pca , scale =0) #plot pca components using biplot in r

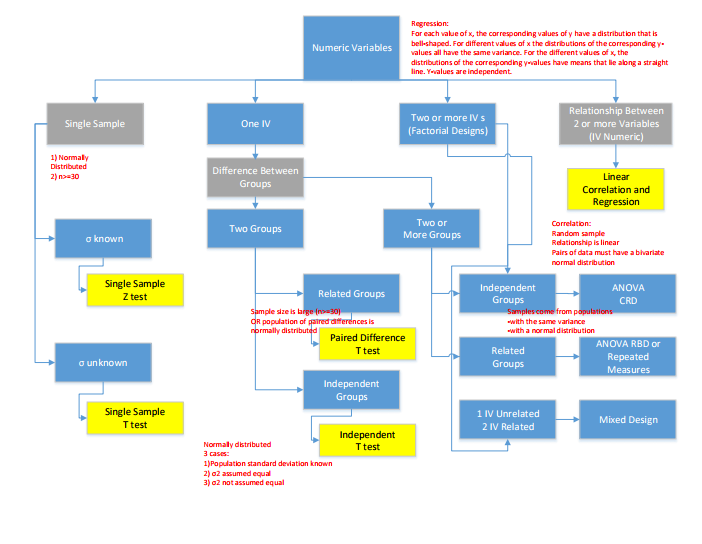




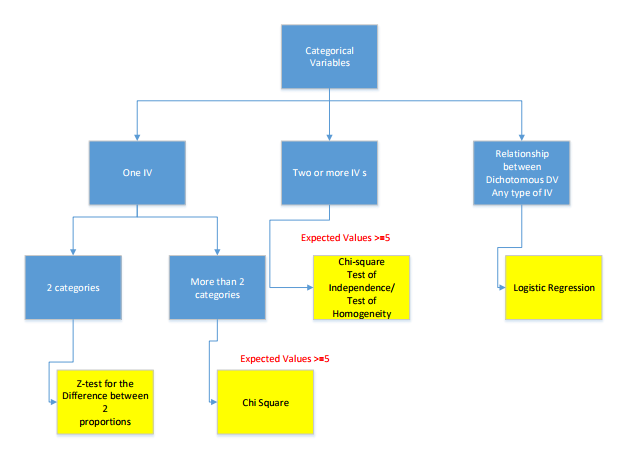
NUM\_CREDIT & JOB are used for predicting DEFAULT

**Hypothesis Test:**

For hypothesis testing we can follow the decision tree diagram below for statistical testing:



For Numerical Variables



For Categorical Variables

We want to see if default will stay below 40% or will increase over 40%. We know the mean for default is 30%.

Null hypothesis: HO *μ* <=.40

Alternate hypothesis: HA *μ* >.40

α =.05

t Stat -6.897204403 < t Critical one-tail 1.646380345

So, we fail to reject HO. The average default won’t go beyond 40% which is good news.



Also, we use ANOVA testing. The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups. We want to see if the means for male single, male divorce and male married or widow are statistically significant different. We use Excel result as below:



Hypothesis:

H0: μ1 = μ2 = μ3

H1: Means are not all equal              α=0.05

F=603.65 > F critical=2.99 so we reject Ho. In fact, the three groups are independent.

**ROI (return on investment):**

I calculate the percentage of defaults (0.3) multiplied by the average loan amount and compared to the original method.  I was then able to show the difference saved by the new method per 1000 loans.  I think showing the ROI is a strong motivator for buy in for the analytic plan and the project as a whole.

count<-table(credit\_data$DEFAULT)

count

percent<-count/1000

percent

barplot(percent, main="Default Distribution", xlab="Loan Default", ylab="%", names.arg = c("Non-default", "Default"), las=1)

*> count*

*0 1*

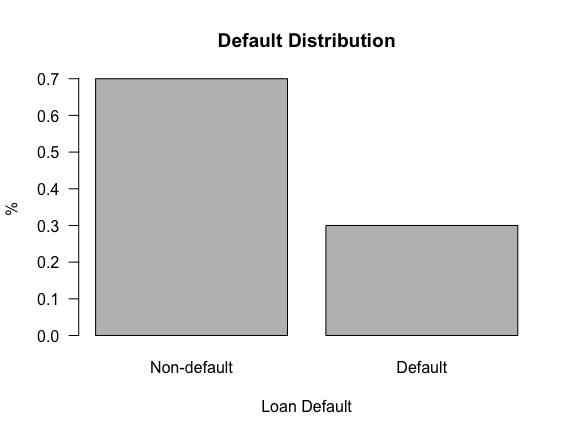
*700 300*

*> percent<-count/1000*

*> percent*

*0 1*

*0.7 0.3*

**

averagedefaultamount<-mean(credit\_data$AMOUNT)\*.3

averagedefaultamount

*[1] 981.3774* # This is average amount loans for default

credit\_data$ROI<-credit\_data$AMOUNT-averagedefaultamount

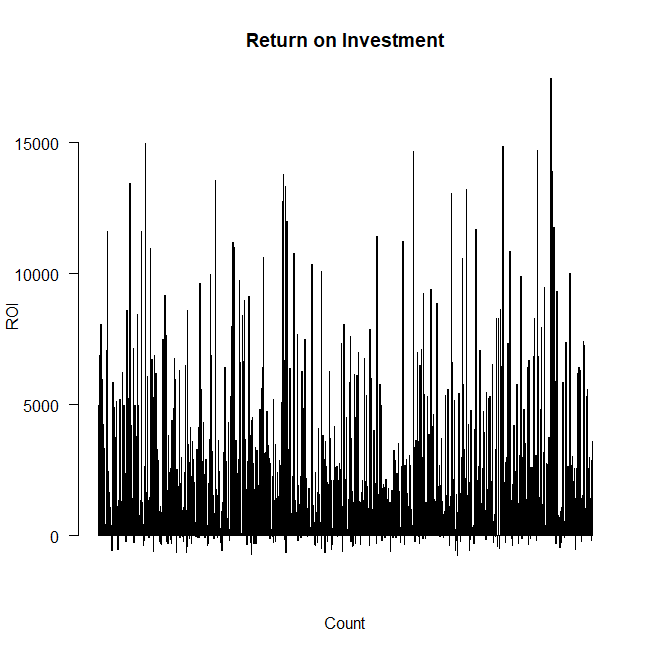
barplot(credit\_data$ROI, main="Return on Investment", xlab="Count", ylab="ROI", las=1)

sum(credit\_data$ROI) #Insert column ROI into the table credit\_data

*[1] 2289881* # This is total return on investment

sum(credit\_data$AMOUNT)

*[1] 3271258* # This is the total amount of all loans



ROI Bar chart

**Logistic Regression:**

We could use linear regression to predict the default using threshold and anything below assigned to 0 and anything above is assigned to 1. Unfortunately, the predicted values could be well outside of the 0 to 1 expected range. Therefore, linear or multivariate regression will not be effective for predicting the values. Instead, logistic regression will be more useful because it will produce probability that the target value is 1. Probabilities are always between 0 and 1 so the output will more closely match the target value range than linear regression.

The model summary shows that the p-values for each coefficient. Alongside these coefficients, the summary gives R’s usual at-a-glance scale of asterisks for significance. Using this scale, we can see that the coefficients for Duration, Furniture, Radio.TV, Retraining, Amount, Employment, Male\_Married or Widow, Present Resident, Age and Job are not significant. We can likely simplify the model by removing these variables and get nearly the same accuracy.

#Logistic regression

credit\_data <- read.csv("~/Documents/SNHU/DAT-650-Q1051 Advanced Data Analytics/Credit\_Data.csv")

logit1<-glm(formula = DEFAULT~CHK\_ACCT + DURATION + HISTORY + NEW\_CAR + USED\_CAR + FURNITURE + RADIO\_TV + EDUCATION + RETRAINING + AMOUNT + SAV\_ACCT + EMPLOYMENT + INSTALL\_RATE + MALE\_DIV + MALE\_SINGLE + MALE\_MAR\_or\_WID + CO\_APPLICANT + GUARANTOR + PRESENT\_RESIDENT + REAL\_ESTATE + PROP\_UNKN\_NONE + AGE + OTHER\_INSTALL + RENT + OWN\_RES + NUM\_CREDITS + JOB + NUM\_DEPENDENTS + TELEPHONE + FOREIGN, family = binomial, data = credit\_data)

logit1

install.packages("caret")

install.packages("e1071")

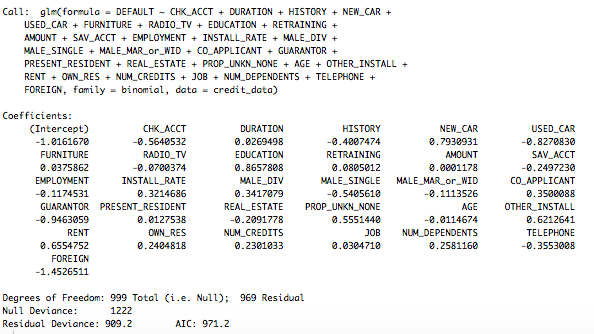
library(caret)

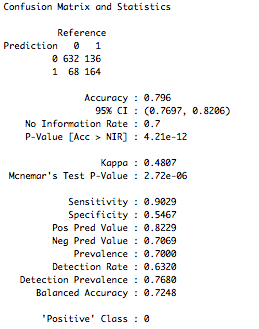
library(e1071)

pdata1 <- predict(logit1, newdata = credit\_data, type = "response")

# use caret and compute a confusion matrix

confusionMatrix(data = as.numeric(pdata1>0.5), reference = credit\_data$DEFAULT)





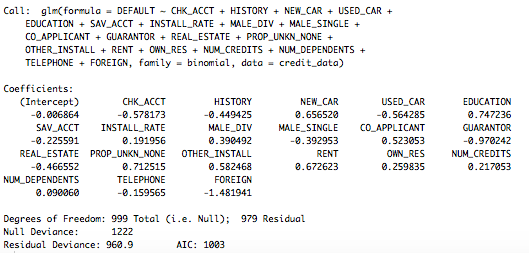
logit2<-glm(formula = DEFAULT~CHK\_ACCT + HISTORY + NEW\_CAR + USED\_CAR + EDUCATION + SAV\_ACCT + INSTALL\_RATE + MALE\_DIV + MALE\_SINGLE + CO\_APPLICANT + GUARANTOR + REAL\_ESTATE + PROP\_UNKN\_NONE + OTHER\_INSTALL + RENT + OWN\_RES + NUM\_CREDITS + NUM\_DEPENDENTS + TELEPHONE + FOREIGN, family = binomial, data = credit\_data)

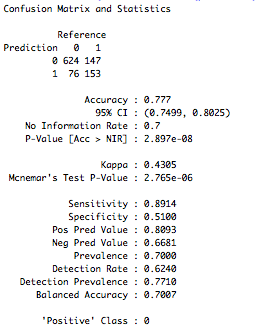
logit2

pdata2 <- predict(logit2, newdata = credit\_data, type = "response")

# use caret and compute a confusion matrix

confusionMatrix(data = as.numeric(pdata2>0.5), reference = credit\_data$DEFAULT)





**PART V: DEPLOYMENT**

The following R code is used for model deployment:

We use packages nnet and pmml to deploy. At the end of this document is the output of model deployment.

#Model deployment

library(nnet)

mydata<- read.csv("~/Documents/SNHU/DAT-650-Q1051 Advanced Data Analytics/Credit\_Data.csv")

mydataNet<-nnet(mydata$DEFAULT~., data=mydata, size=4)

install.packages("pmml")

library(pmml)

pmml(mydataNet)

**PART VI: CONCLUSION**

In this assignment, I have demonstrated mastery of the following course outcomes:

• Defend the value and purpose of data collection and analytics structures for institutional and organizational progress

• Evaluate data analytic architectures for potential security, privacy, and ethical concerns for identification of software solutions

• Create models within various environments by assessing the applicability and value of data strategies

• Create pilot data analytic solution stack plans that address identified organizational data issues

• Present proposals for full implementation of data analytic solution stacks based on identified data needs and pilot outcomes

I have developed a data analysis enterprise strategy that will add business value. Business value, in this assessment, is the value or benefit added to the organization from the incorporation and use of a data strategy. I developed an architecture and strategy, and then convinced the organizational executives that my plan will add necessary value to the organization.

References:

The Data Analyst Role. Retrieved from <http://www.modernanalyst.com/TheProfession/Roles/DataAnalyst/tabid/189/Default.aspx>

[Why the R Programming Language Is Good For Business](https://www.fastcompany.com/3030063/why-the-r-programming-language-is-good-for-business), 5/5/2014. Retrieved from

<https://www.fastcompany.com/3030063/why-the-r-programming-language-is-good-for-business>

Master Data Management and Data Governance, Alex Berson, Larry Dubov, 2nd Edition, Mc Graw Hill.

Business Management: Ethical Issues in Finance. Retrieved from: <https://www.moneyinstructor.com/doc/ethicalfinance.asp>

Driving growth in 2017 with the help of credit & finance. Source: [D&B](http://www.dnb.com/perspectives/finance-credit-risk/how-credit-finance-teams-can-drive-growth-2017.html).

Retrieved from: http://blog.bisnode.hu/bisnode/driving-growth-2017-help-credit-finance/

Credit risk. 1982 - 2017 by [William C. Spaulding](http://thismatter.com/about.htm). Retrieved from:

<http://thismatter.com/money/banking/bank-risks.htm>

[Preventing & resolving problem loans - GBRW Consulting](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=8&cad=rja&uact=8&ved=0ahUKEwjms-Sd55bXAhXCSyYKHf-IBUwQFghUMAc&url=http%3A%2F%2Fwww.gbrw.com%2F_literature_129465%2FPreventing_and_Resolving_Non-Performing_Loans&usg=AOvVaw02svnO0DK3DBgnxNWjASgM). Retrieved from:

[www.gbrw.com/\_literature.../Preventing\_and\_Resolving\_Non-Performing\_Loans](http://www.gbrw.com/_literature.../Preventing_and_Resolving_Non-Performing_Loans)

Selling Analytics to your Stakeholders, by [Jonathan Solórzano-Hamilton](https://towardsdatascience.com/@peachpie?source=post_header_lockup). Retrieved from:

<https://towardsdatascience.com/selling-analytics-to-your-stakeholders-3e516cd6cc2b>

How lenders decide whether to give you credit. Retrieved from:

<https://www.citizensadvice.org.uk/debt-and-money/borrowing-money/how-lenders-decide-whether-to-give-you-credit/>

Subject Matter Expert. Retrieved from:

<http://epf.eclipse.org/wikis/epfpractices/core.default.role_def.base/roles/subject_matter_expert_D73B48C5.html>

Stakeholder Governance: How Stakeholders Influence Corporate Decision Making. Retrieved from <https://www.researchgate.net/publication/228221781_Stakeholder_Governance_How_Stakeholders_Influence_Corporate_Decision_Making>

Analysis of Credit Approval Data, Ryan Kuhn. Retrieved from: http://rstudio-pubs-static.s3.amazonaws.com/73039\_9946de135c0a49daa7a0a9eda4a67a72.html

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**APPENDIX:**

**Model Deployment output:**

*<PMML version="4.3" xmlns="http://www.dmg.org/PMML-4\_3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.dmg.org/PMML-4\_3 http://www.dmg.org/pmml/v4-3/pmml-4-3.xsd">*

*<Header copyright="Copyright (c) 2018 hhuynh" description="Neural Network PMML Model">*

*<Extension name="user" value="hhuynh" extender="Rattle/PMML"/>*

*<Application name="Rattle/PMML" version="1.4"/>*

*<Timestamp>2018-01-20 17:47:51</Timestamp>*

*</Header>*

*<DataDictionary numberOfFields="32">*

*<DataField name="mydata$DEFAULT" optype="continuous" dataType="double"/>*

*<DataField name="OBS." optype="continuous" dataType="double"/>*

*<DataField name="CHK\_ACCT" optype="continuous" dataType="double"/>*

*<DataField name="DURATION" optype="continuous" dataType="double"/>*

*<DataField name="HISTORY" optype="continuous" dataType="double"/>*

*<DataField name="NEW\_CAR" optype="continuous" dataType="double"/>*

*<DataField name="USED\_CAR" optype="continuous" dataType="double"/>*

*<DataField name="FURNITURE" optype="continuous" dataType="double"/>*

*<DataField name="RADIO\_TV" optype="continuous" dataType="double"/>*

*<DataField name="EDUCATION" optype="continuous" dataType="double"/>*

*<DataField name="RETRAINING" optype="continuous" dataType="double"/>*

*<DataField name="AMOUNT" optype="continuous" dataType="double"/>*

*<DataField name="SAV\_ACCT" optype="continuous" dataType="double"/>*

*<DataField name="EMPLOYMENT" optype="continuous" dataType="double"/>*

*<DataField name="INSTALL\_RATE" optype="continuous" dataType="double"/>*

*<DataField name="MALE\_DIV" optype="continuous" dataType="double"/>*

*<DataField name="MALE\_SINGLE" optype="continuous" dataType="double"/>*

*<DataField name="MALE\_MAR\_or\_WID" optype="continuous" dataType="double"/>*

*<DataField name="CO\_APPLICANT" optype="continuous" dataType="double"/>*

*<DataField name="GUARANTOR" optype="continuous" dataType="double"/>*

*<DataField name="PRESENT\_RESIDENT" optype="continuous" dataType="double"/>*

*<DataField name="REAL\_ESTATE" optype="continuous" dataType="double"/>*

*<DataField name="PROP\_UNKN\_NONE" optype="continuous" dataType="double"/>*

*<DataField name="AGE" optype="continuous" dataType="double"/>*

*<DataField name="OTHER\_INSTALL" optype="continuous" dataType="double"/>*

*<DataField name="RENT" optype="continuous" dataType="double"/>*

*<DataField name="OWN\_RES" optype="continuous" dataType="double"/>*

*<DataField name="NUM\_CREDITS" optype="continuous" dataType="double"/>*

*<DataField name="JOB" optype="continuous" dataType="double"/>*

*<DataField name="NUM\_DEPENDENTS" optype="continuous" dataType="double"/>*

*<DataField name="TELEPHONE" optype="continuous" dataType="double"/>*

*<DataField name="FOREIGN" optype="continuous" dataType="double"/>*

*</DataDictionary>*

*<NeuralNetwork modelName="NeuralNet\_model" functionName="regression" numberOfLayers="2" activationFunction="logistic">*

*<MiningSchema>*

*<MiningField name="mydata$DEFAULT" usageType="predicted" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="OBS." usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="CHK\_ACCT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="DURATION" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="HISTORY" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="NEW\_CAR" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="USED\_CAR" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="FURNITURE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="RADIO\_TV" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="EDUCATION" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="RETRAINING" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="AMOUNT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="SAV\_ACCT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="EMPLOYMENT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="INSTALL\_RATE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="MALE\_DIV" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="MALE\_SINGLE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="MALE\_MAR\_or\_WID" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="CO\_APPLICANT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="GUARANTOR" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="PRESENT\_RESIDENT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="REAL\_ESTATE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="PROP\_UNKN\_NONE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="AGE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="OTHER\_INSTALL" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="RENT" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="OWN\_RES" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="NUM\_CREDITS" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="JOB" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="NUM\_DEPENDENTS" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="TELEPHONE" usageType="active" invalidValueTreatment="returnInvalid"/>*

*<MiningField name="FOREIGN" usageType="active" invalidValueTreatment="returnInvalid"/>*

*</MiningSchema>*

*<Output>*

*<OutputField name="Predicted\_mydata$DEFAULT" feature="predictedValue"/>*

*</Output>*

*<NeuralInputs numberOfInputs="31">*

*<NeuralInput id="1">*

*<DerivedField name="derivedNI\_OBS." optype="continuous" dataType="double">*

*<FieldRef field="OBS."/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="2">*

*<DerivedField name="derivedNI\_CHK\_ACCT" optype="continuous" dataType="double">*

*<FieldRef field="CHK\_ACCT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="3">*

*<DerivedField name="derivedNI\_DURATION" optype="continuous" dataType="double">*

*<FieldRef field="DURATION"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="4">*

*<DerivedField name="derivedNI\_HISTORY" optype="continuous" dataType="double">*

*<FieldRef field="HISTORY"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="5">*

*<DerivedField name="derivedNI\_NEW\_CAR" optype="continuous" dataType="double">*

*<FieldRef field="NEW\_CAR"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="6">*

*<DerivedField name="derivedNI\_USED\_CAR" optype="continuous" dataType="double">*

*<FieldRef field="USED\_CAR"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="7">*

*<DerivedField name="derivedNI\_FURNITURE" optype="continuous" dataType="double">*

*<FieldRef field="FURNITURE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="8">*

*<DerivedField name="derivedNI\_RADIO\_TV" optype="continuous" dataType="double">*

*<FieldRef field="RADIO\_TV"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="9">*

*<DerivedField name="derivedNI\_EDUCATION" optype="continuous" dataType="double">*

*<FieldRef field="EDUCATION"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="10">*

*<DerivedField name="derivedNI\_RETRAINING" optype="continuous" dataType="double">*

*<FieldRef field="RETRAINING"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="11">*

*<DerivedField name="derivedNI\_AMOUNT" optype="continuous" dataType="double">*

*<FieldRef field="AMOUNT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="12">*

*<DerivedField name="derivedNI\_SAV\_ACCT" optype="continuous" dataType="double">*

*<FieldRef field="SAV\_ACCT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="13">*

*<DerivedField name="derivedNI\_EMPLOYMENT" optype="continuous" dataType="double">*

*<FieldRef field="EMPLOYMENT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="14">*

*<DerivedField name="derivedNI\_INSTALL\_RATE" optype="continuous" dataType="double">*

*<FieldRef field="INSTALL\_RATE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="15">*

*<DerivedField name="derivedNI\_MALE\_DIV" optype="continuous" dataType="double">*

*<FieldRef field="MALE\_DIV"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="16">*

*<DerivedField name="derivedNI\_MALE\_SINGLE" optype="continuous" dataType="double">*

*<FieldRef field="MALE\_SINGLE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="17">*

*<DerivedField name="derivedNI\_MALE\_MAR\_or\_WID" optype="continuous" dataType="double">*

*<FieldRef field="MALE\_MAR\_or\_WID"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="18">*

*<DerivedField name="derivedNI\_CO\_APPLICANT" optype="continuous" dataType="double">*

*<FieldRef field="CO\_APPLICANT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="19">*

*<DerivedField name="derivedNI\_GUARANTOR" optype="continuous" dataType="double">*

*<FieldRef field="GUARANTOR"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="20">*

*<DerivedField name="derivedNI\_PRESENT\_RESIDENT" optype="continuous" dataType="double">*

*<FieldRef field="PRESENT\_RESIDENT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="21">*

*<DerivedField name="derivedNI\_REAL\_ESTATE" optype="continuous" dataType="double">*

*<FieldRef field="REAL\_ESTATE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="22">*

*<DerivedField name="derivedNI\_PROP\_UNKN\_NONE" optype="continuous" dataType="double">*

*<FieldRef field="PROP\_UNKN\_NONE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="23">*

*<DerivedField name="derivedNI\_AGE" optype="continuous" dataType="double">*

*<FieldRef field="AGE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="24">*

*<DerivedField name="derivedNI\_OTHER\_INSTALL" optype="continuous" dataType="double">*

*<FieldRef field="OTHER\_INSTALL"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="25">*

*<DerivedField name="derivedNI\_RENT" optype="continuous" dataType="double">*

*<FieldRef field="RENT"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="26">*

*<DerivedField name="derivedNI\_OWN\_RES" optype="continuous" dataType="double">*

*<FieldRef field="OWN\_RES"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="27">*

*<DerivedField name="derivedNI\_NUM\_CREDITS" optype="continuous" dataType="double">*

*<FieldRef field="NUM\_CREDITS"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="28">*

*<DerivedField name="derivedNI\_JOB" optype="continuous" dataType="double">*

*<FieldRef field="JOB"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="29">*

*<DerivedField name="derivedNI\_NUM\_DEPENDENTS" optype="continuous" dataType="double">*

*<FieldRef field="NUM\_DEPENDENTS"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="30">*

*<DerivedField name="derivedNI\_TELEPHONE" optype="continuous" dataType="double">*

*<FieldRef field="TELEPHONE"/>*

*</DerivedField>*

*</NeuralInput>*

*<NeuralInput id="31">*

*<DerivedField name="derivedNI\_FOREIGN" optype="continuous" dataType="double">*

*<FieldRef field="FOREIGN"/>*

*</DerivedField>*

*</NeuralInput>*

*</NeuralInputs>*

*<NeuralLayer numberOfNeurons="4">*

*<Neuron id="32" bias="-0.193870865111441">*

*<Con from="1" weight="-6.65204328776498"/>*

*<Con from="2" weight="0.212161236399575"/>*

*<Con from="3" weight="-0.625727420477847"/>*

*<Con from="4" weight="-0.00834626047482282"/>*

*<Con from="5" weight="-0.549524342412703"/>*

*<Con from="6" weight="0.543465734133497"/>*

*<Con from="7" weight="0.175478160764934"/>*

*<Con from="8" weight="-0.375552136080713"/>*

*<Con from="9" weight="-0.227729299385101"/>*

*<Con from="10" weight="-0.42143891309388"/>*

*<Con from="11" weight="-5.7137467470756"/>*

*<Con from="12" weight="-0.56919942126039"/>*

*<Con from="13" weight="0.221900058733498"/>*

*<Con from="14" weight="0.30954402371831"/>*

*<Con from="15" weight="0.450631299754605"/>*

*<Con from="16" weight="-0.52130878440159"/>*

*<Con from="17" weight="0.227460411041764"/>*

*<Con from="18" weight="0.204671313520521"/>*

*<Con from="19" weight="0.5876946725266"/>*

*<Con from="20" weight="-0.654756521163767"/>*

*<Con from="21" weight="-0.344837320662238"/>*

*<Con from="22" weight="-0.355897703487426"/>*

*<Con from="23" weight="0.235042162225863"/>*

*<Con from="24" weight="0.629708848516814"/>*

*<Con from="25" weight="0.0591563828400389"/>*

*<Con from="26" weight="-0.142779098625408"/>*

*<Con from="27" weight="-0.0331879406979615"/>*

*<Con from="28" weight="-0.62344168683104"/>*

*<Con from="29" weight="-0.632107590082869"/>*

*<Con from="30" weight="0.0192399220498122"/>*

*<Con from="31" weight="0.247622185563458"/>*

*</Neuron>*

*<Neuron id="33" bias="-0.389673965630903">*

*<Con from="1" weight="0.63775439811694"/>*

*<Con from="2" weight="-0.0839118784411424"/>*

*<Con from="3" weight="0.124406020356226"/>*

*<Con from="4" weight="0.388587333515691"/>*

*<Con from="5" weight="-0.308603474637493"/>*

*<Con from="6" weight="-0.446043455367908"/>*

*<Con from="7" weight="-0.498287573317066"/>*

*<Con from="8" weight="-0.649936928308908"/>*

*<Con from="9" weight="0.120303076194156"/>*

*<Con from="10" weight="0.562180921807885"/>*

*<Con from="11" weight="-0.511911230537261"/>*

*<Con from="12" weight="-0.0951889390623455"/>*

*<Con from="13" weight="0.469037581238367"/>*

*<Con from="14" weight="0.495595997749664"/>*

*<Con from="15" weight="0.504186458094045"/>*

*<Con from="16" weight="0.104324071854353"/>*

*<Con from="17" weight="-0.362962712979114"/>*

*<Con from="18" weight="-0.549367280071601"/>*

*<Con from="19" weight="0.295198523878881"/>*

*<Con from="20" weight="0.310243819857908"/>*

*<Con from="21" weight="-0.487171795207414"/>*

*<Con from="22" weight="0.536339192232117"/>*

*<Con from="23" weight="0.190708458936737"/>*

*<Con from="24" weight="-0.520946748613062"/>*

*<Con from="25" weight="-0.291936983898725"/>*

*<Con from="26" weight="-0.0934996814349328"/>*

*<Con from="27" weight="-0.276205521495438"/>*

*<Con from="28" weight="-0.332383930456187"/>*

*<Con from="29" weight="-0.404006140312474"/>*

*<Con from="30" weight="0.297379933195951"/>*

*<Con from="31" weight="0.266784011758864"/>*

*</Neuron>*

*<Neuron id="34" bias="0.637325845658779">*

*<Con from="1" weight="-0.311417069053277"/>*

*<Con from="2" weight="-0.581379299936816"/>*

*<Con from="3" weight="0.448509921133518"/>*

*<Con from="4" weight="-0.464246333995834"/>*

*<Con from="5" weight="-0.631333485804498"/>*

*<Con from="6" weight="0.155970278661698"/>*

*<Con from="7" weight="0.267563276272267"/>*

*<Con from="8" weight="-0.373336080135778"/>*

*<Con from="9" weight="-0.412173624942079"/>*

*<Con from="10" weight="-0.168358926381916"/>*

*<Con from="11" weight="-0.632987546082586"/>*

*<Con from="12" weight="-0.610790718020871"/>*

*<Con from="13" weight="0.00729702604003246"/>*

*<Con from="14" weight="0.135445563355461"/>*

*<Con from="15" weight="-0.245741701265797"/>*

*<Con from="16" weight="0.447858253819868"/>*

*<Con from="17" weight="0.376294414140284"/>*

*<Con from="18" weight="-0.532482823589817"/>*

*<Con from="19" weight="-0.615217917459085"/>*

*<Con from="20" weight="-0.453598279925063"/>*

*<Con from="21" weight="0.153761152783409"/>*

*<Con from="22" weight="-0.032414482999593"/>*

*<Con from="23" weight="0.223409404000267"/>*

*<Con from="24" weight="0.317792993364856"/>*

*<Con from="25" weight="0.63004810065031"/>*

*<Con from="26" weight="-0.560800576163456"/>*

*<Con from="27" weight="0.693177536176518"/>*

*<Con from="28" weight="-0.207406022306532"/>*

*<Con from="29" weight="-0.563223484810442"/>*

*<Con from="30" weight="0.405437423940748"/>*

*<Con from="31" weight="-0.353130237059668"/>*

*</Neuron>*

*<Neuron id="35" bias="-0.438092007052994">*

*<Con from="1" weight="-33.6834543099268"/>*

*<Con from="2" weight="0.445020951421268"/>*

*<Con from="3" weight="-0.345381745874346"/>*

*<Con from="4" weight="-0.338561056532769"/>*

*<Con from="5" weight="-0.325584357245067"/>*

*<Con from="6" weight="-0.387967556994408"/>*

*<Con from="7" weight="-0.609805508516729"/>*

*<Con from="8" weight="0.348370197069398"/>*

*<Con from="9" weight="-0.348055831573207"/>*

*<Con from="10" weight="-0.387405088124797"/>*

*<Con from="11" weight="-30.7898684603603"/>*

*<Con from="12" weight="0.0587534873249724"/>*

*<Con from="13" weight="0.00309189055206194"/>*

*<Con from="14" weight="-0.292170370492722"/>*

*<Con from="15" weight="-0.0651560267433524"/>*

*<Con from="16" weight="-0.551322546391428"/>*

*<Con from="17" weight="0.239353464997456"/>*

*<Con from="18" weight="0.142221155017614"/>*

*<Con from="19" weight="0.563665698142722"/>*

*<Con from="20" weight="0.380272332691483"/>*

*<Con from="21" weight="0.447826946229217"/>*

*<Con from="22" weight="-0.541309855598956"/>*

*<Con from="23" weight="-0.992071603333258"/>*

*<Con from="24" weight="-0.289684486967945"/>*

*<Con from="25" weight="0.408132049304122"/>*

*<Con from="26" weight="0.451780014545127"/>*

*<Con from="27" weight="-0.227359505809117"/>*

*<Con from="28" weight="-0.507534135151372"/>*

*<Con from="29" weight="-0.0143263251361464"/>*

*<Con from="30" weight="0.429153391917196"/>*

*<Con from="31" weight="-0.0326612822245806"/>*

*</Neuron>*

*</NeuralLayer>*

*<NeuralLayer numberOfNeurons="1">*

*<Neuron id="36" bias="-161.540168827493">*

*<Con from="32" weight="-2.06005359666553"/>*

*<Con from="33" weight="-0.550748661900424"/>*

*<Con from="34" weight="-0.256506304675713"/>*

*<Con from="35" weight="-3.56838570853035"/>*

*</Neuron>*

*</NeuralLayer>*

*<NeuralOutputs numberOfOutputs="1">*

*<NeuralOutput outputNeuron="36">*

*<DerivedField name="derivedNO\_mydata$DEFAULT" optype="continuous" dataType="double">*

*<FieldRef field="mydata$DEFAULT"/>*

*</DerivedField>*

*</NeuralOutput>*

*</NeuralOutputs>*

*</NeuralNetwork>*

*</PMML>*