Analysis of Simulated Loan Data using Clustering Analysis and Graphical Modeling for Prediction of Default Classification and Variable Analysis

Ryan Funderburk (rpf2fh)

Abstract: A major problem faced by banks and lending institutions is the loss of cash and assets through borrowers defaulting on loans. The goal of this analysis is to understand how demographic variables of loan applicants and variables relating to the structure of loans contribute to the applicants' probabilities of defaulting. The data used in this analysis is simulated based on data from the Credit Bureau. The steps of analysis in this paper include dimensional reduction through Principal Component Analysis to understand the variation explained through each Principal component and the variables with the largest scores in each component. Next, K-means and Gaussian Mixture Model clustering was applied to the first two Principal Components to classify observations into clusters based on default or not defaulting. Finally, a Bayesian Network model is learned on the loan variables using Min-Max Hill Climbing to understand the relationships and joint conditional distributions of the variables, and the model is used to predict probability estimates of an applicant defaulting are obtained given values for each node of the Bayesian Network.

Initial Motivation and Exploration of the Data:

Data:

The data is Credit Risk data on loans and applicants that has been simulated based on Credit Bureau Data. The data set was uploaded to Kaggle.com by Lao Tse as “Credit Risk Dataset”, and the method behind the simulation of the data is not specified. The variables simulated are demographic data, loan structure variables, and credit history of loan applicants. This data set contains over 30,000 observations of the 12 variables. The variable of interest is *loan_status*, which gives information on the default status of the applicant after the term of the loan. Notable characteristics of the data are that the response is binary, and the explanatory variables consist of 4 categorical variables and 7 continuous variables. The demographic information includes age, income, housing status, and years employed. The loan data includes the amount of the loan, the interest rate, the grade, purpose, and default history of the borrowers.

Motivation:

This data should be studied more closely because credit is a major part of how customers shop, and a major source of economic power for buyers. Analysis of Credit Risk data can help institutions to understand the risk factors associated with defaulted applicants and how specific variables can contribute to one's credit risk. Understanding this data may allow lending institutions to improve their lending without incurring losses and help consumers to understand how they can improve their applications for more borrowing. This dataset specifically can help lenders create models and methods that can predict the probability of default based on an applicant's demographic data and the specifics of their loan agreements in order to improve their lending process.

Problem:

The questions I intend to answer through the analysis of this data are:

1. Which demographic and loan structure variables contribute to the most variation in the data set, and can these variables be used to predict/classify the status of a loan?
2. What is the structure of the relationship between an applicant's loan variables and default status, and can the likelihood of an applicant defaulting be predicted given specific loan structures?

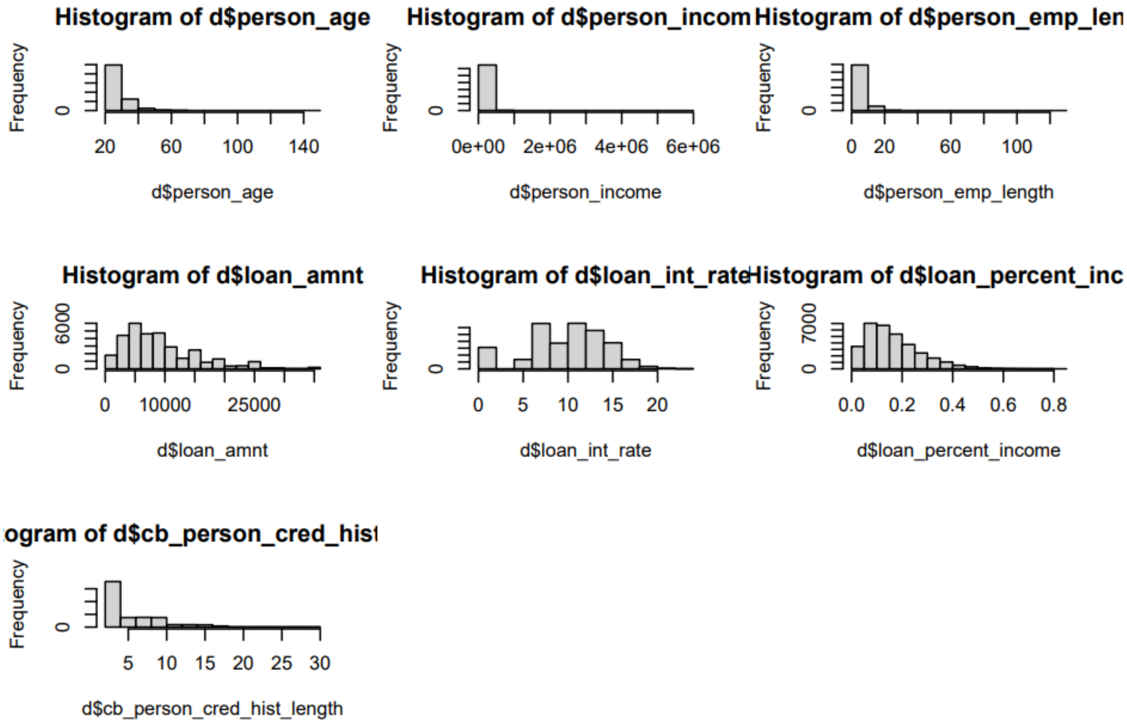
This data set has potential problems that may concern the results of the analysis and the predictability of the models created. The first concern is that the data set isn’t representative of a real world sample since it has been simulated based on Credit Bureau Data. This brings up concern for the applicability of the models created through this data to problems and applications for real world lending institutions. Due to the nature of privacy laws and private institutions, there is a lack of real world credit risk data available for analysis that hasn’t been extensively researched and studied, so this data set will suffice for analysis so long as the simulation produces values for observations that are logical given existing knowledge of Credit

Bureau data. The next concern that arises for this data set is the existence of several outliers and illogical values for specific variables, and to address this concern those data points will be removed before analysis. The final problem for this data set is the interest rate variable consists of several missing values. This issue was addressed by removing the observations with missing values, and it shouldn't be of concern that these values were removed because the data set is sufficiently large enough for the methods used without them.

<u>Preliminary Work:</u>

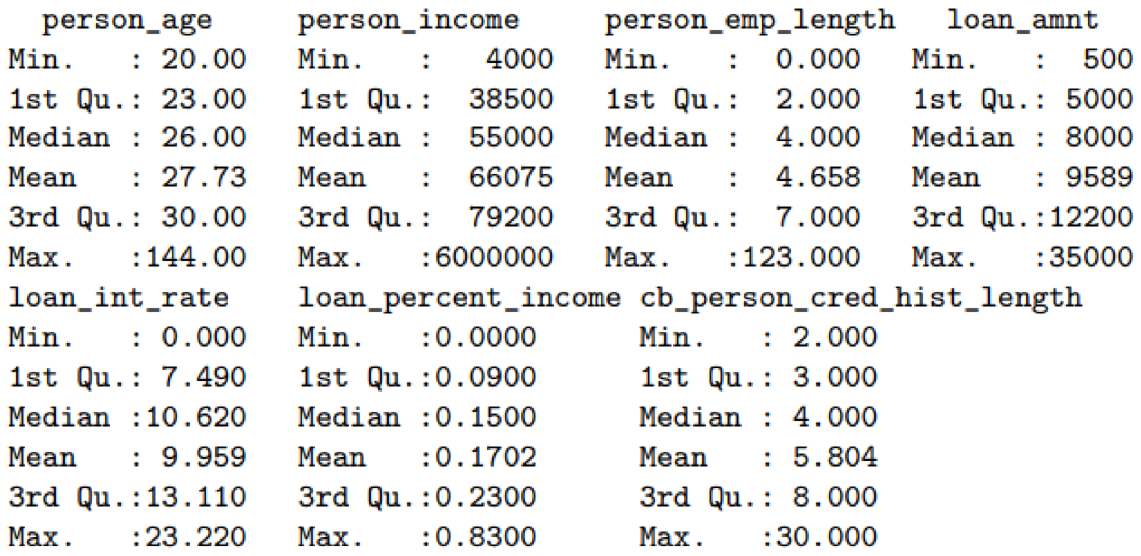
In previous work analyzing data sets of credit cards and loan agreements, logistic regression has been used to model factors and try to predict default probabilities. One previous study on small business lending assessing credit risk found models through probabilistic neural networks had lower Type 1 errors than logistic regression decision trees. A study found a fuzzy model based on using a set of uncertainty procedures was best when predicting the number of non-performing loans using data from Iranian Banks.

<u>Exploratory Data Analysis:</u>



The histograms of the quantitative variables show mostly right-skewed distributions. The *loan_amnt*, *loan_int_rate*, and *loan_percent_inc* variables display almost normal distributions that are slightly skewed by outliers that can be removed. The variables *person_age* and *person_emp_length* show a concerning outlier for a variable with a value above one hundred, which is unreasonable given the basic understanding of human life span.

Notably, *person_age* exhibited a max value of 144 and *person_emp_length* exhibited a max of 123. These outliers are further exemplified in the five number summaries of the quantitative variables listed here:

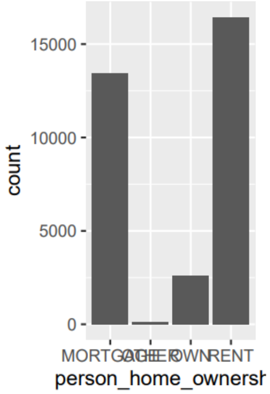


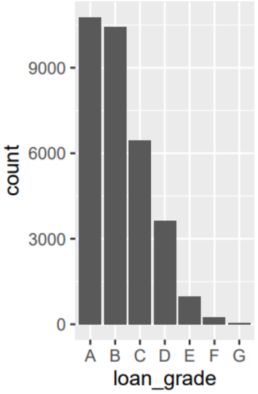
```
  person_age      person_income     person_emp_length   loan_amnt    
 Min.   : 20.00   Min.   :   4000   Min.   :  0.000   Min.   :  500  
 1st Qu.: 23.00   1st Qu.:  38500   1st Qu.:  2.000   1st Qu.: 5000  
 Median : 26.00   Median :  55000   Median :  4.000   Median : 8000  
 Mean   : 27.73   Mean   :  66075   Mean   :  4.658   Mean   : 9589  
 3rd Qu.: 30.00   3rd Qu.:  79200   3rd Qu.:  7.000   3rd Qu.:12200  
 Max.   :144.00   Max.   :6000000   Max.   :123.000   Max.   :35000  
 loan_int_rate    loan_percent_income cb_person_cred_hist_length
 Min.   : 0.000   Min.   :0.0000      Min.   : 2.000            
 1st Qu.: 7.490   1st Qu.:0.0900      1st Qu.: 3.000            
 Median :10.620   Median :0.1500      Median : 4.000            
 Mean   : 9.959   Mean   :0.1702      Mean   : 5.804            
 3rd Qu.:13.110   3rd Qu.:0.2300      3rd Qu.: 8.000            
 Max.   :23.220   Max.   :0.8300      Max.   :30.000            
```

The response variable *loan_status* isn't balanced in its responses, the value for a default or "1" is only 21.8% of the observations and the value for non-defaulted loans or "0" is 78.2% of the observations in the data set. This may cause our methods for classification to predict more false negative values on the test data because our model is trained on data that is favorably balanced towards zero values.

The Pearson correlation coefficients for *person_age* and *person_emp_length*, *loan_amnt* and *loan_percent_inc*, and *person_income* and *loan_percent_inc* were notably high. There may be some multicollinearity present between these variables, and this makes sense given the definition and collection of these variables. Higher values of age might imply that someone has worked longer because they would be of working age for longer, and the loan percent of income is calculated using income and loan amount variables. These relationships will be taken into consideration when interpreting the results of analysis and applying models to the data.

Analyzing the histograms of the categorical variables of this data set shows some interesting insights. The histogram of *loan_grade* shows that most loans are graded A or B versus lower grades of C-G. Next, the histogram of *person_home_ownership* shows that most homes are mortgaged or rented by the applicants in this data set.





Analysis of Question 1:

Motivation

This data is important for lenders to understand its patterns and trends, so they can distribute credit safely to borrowers while minimizing their potential losses to defaults. Understanding how independent factors of borrowers affect their chances at defaulting can help lenders to better distribute credit to people who are most likely to pay back and make good on their debt.

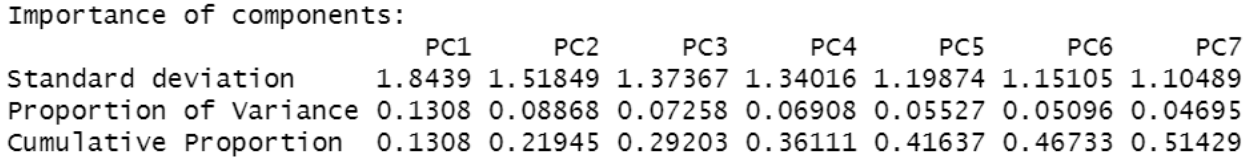
Methodology

First, this data set has high dimensions, especially categorical variables using one-hot encoding on dummy variables. The data set has 12 dimensions without dummy coding, and 26 dimensions when dummy coding. First, Principal Component Analysis is applied to reduce the dimensionality of the data, so that clustering analysis can be applied to the first two Principal Components for improved accuracy and clustering results. Then, K-means and Gaussian Mixture Modeling clustering methods are run on the components to attempt to find meaningful clusters of borrower types, and comparisons are run on the distributions of the individual clusters of borrowers. In hyperparameter selection for the clustering methods, k=2 clusters for K-means and 2 Gaussian Components for GMM were selected to attempt to classify applicants on loan status by cluster. This hyperparameter selection directly clashed with the results of the elbow method, which suggested 3-means clustering, but since the goal is the prediction of a default it makes more sense to cluster the data into 2 groups one for default and one for non-default. In order to explore how the factors of a loan relate to default risk, PCA will allow for the dimensionality of the data set to be reduced in order to run logistic regression on the dataset removing any concerns of multicollinearity.

Results:

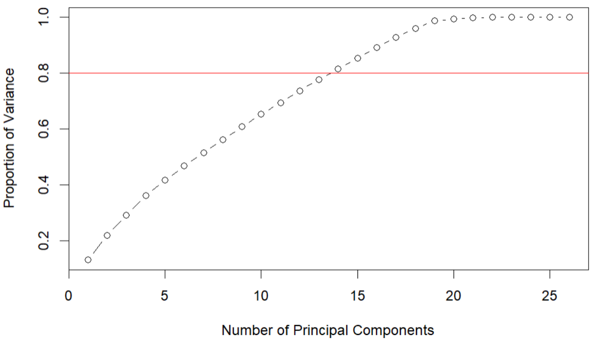
PCA:

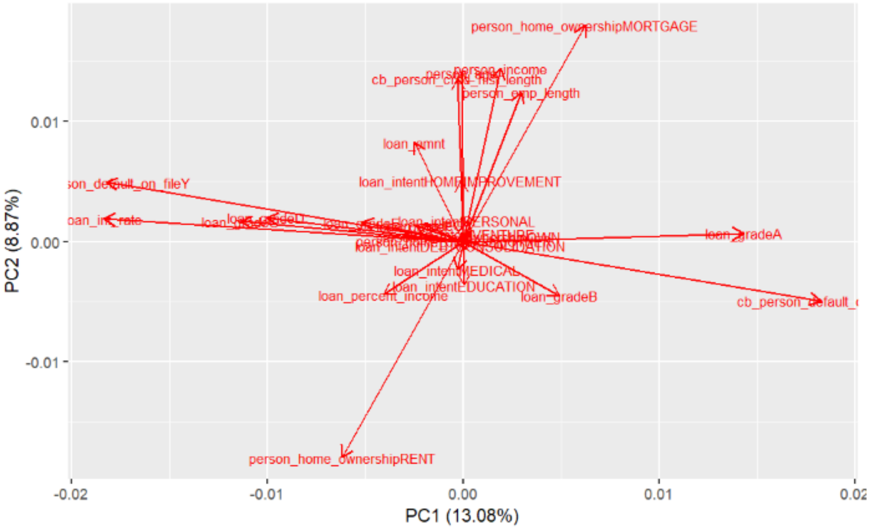
PCA on the entire dataset of 26 variables, removing the variable *loan_status*, as it is the variable of interest that will be used to check the results of clustering. The proportion of variance explained by each of the first seven of the twenty-six principal components is listed:



```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.8439 1.51849 1.37367 1.34016 1.19874 1.15105 1.10489
Proportion of Variance 0.1308 0.08868 0.07258 0.06908 0.05527 0.05096 0.04695
Cumulative Proportion  0.1308 0.21945 0.29203 0.36111 0.41637 0.46733 0.51429
```

The first seven components explain approximately 51.4% of the variance, and the proportion explained by the first principal component is 13.1%. In taking the general rule of thumb of around 80% proportion of variation, the analysis selected to move forward with the first 14 principal components, reducing the dimensionality of our data by 12 dimensions.



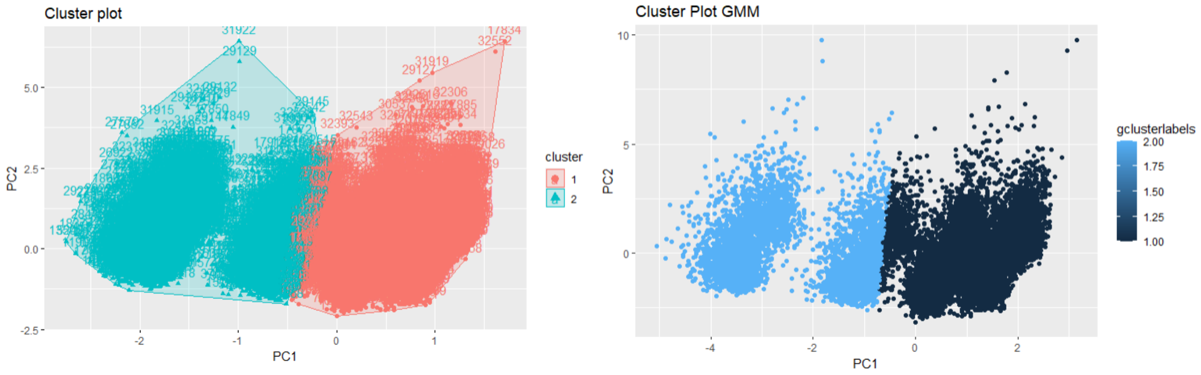


According to the loading plot of the first two components, the variables for default on file yes and no, the interest rate, and loan grade of A have the largest effects on the variability of the first principal component. The variables ownership rent and mortgage, have the highest effect on the second principal component.

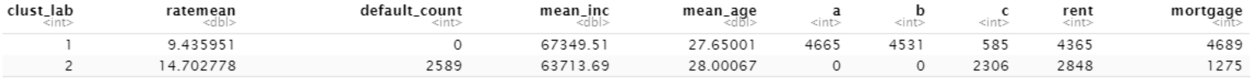
K-Means vs. Gaussian Mixture Modeling on first two Principal Components:

After the dimensionality of the data has been reduced, I moved to applying K-means clustering and Gaussian Mixture Modeling to the first two principal components. I clustered the principal components using 2 centers. In comparing the cluster labels created through K-means to the actual labels of the *loan_status* variable, K-means correctly classifies 10442 out of the 14312 variables at an accuracy rate of 73% on the training data. The clustering labels of the Gaussian Mixture Model correctly predicted 10410 out of the 14312 *loan_status* labels on the training data for a classification accuracy rate of 72.7%.

Cluster plot for K-means (left) and Gaussian Mixture Modeling (right):

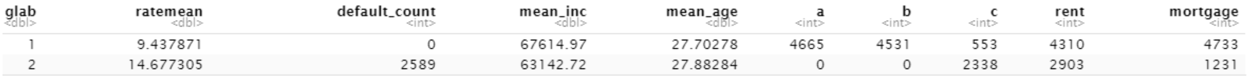


Summaries of Variables in each Cluster (K-means):



clust_lab <int>	ratemean <dbl>	default_count <int>	mean_inc <dbl>	mean_age <dbl>	a <int>	b <int>	c <int>	rent <int>	mortgage <int>
1	9.435951	0	67349.51	27.65001	4665	4531	585	4365	4689
2	14.702778	2589	63713.69	28.00067	0	0	2306	2848	1275

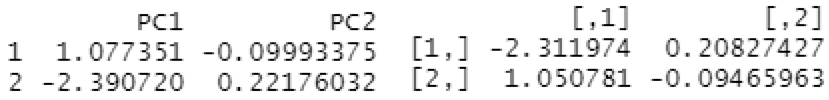
Summaries of Variables in each Cluster (Gaussian Mixture Modeling):



glab <dbl>	ratemean <dbl>	default_count <int>	mean_inc <dbl>	mean_age <dbl>	a <int>	b <int>	c <int>	rent <int>	mortgage <int>
1	9.437871	0	67614.97	27.70278	4665	4531	553	4310	4733
2	14.677305	2589	63142.72	27.88284	0	0	2338	2903	1231

The values of the means and counts of these variables were not that surprising. They make sense in terms of the structure and previous knowledge of loans. They especially line up with the loadings of the PCA components. It makes sense that higher interest rates would lead to more defaults, previous defaults would suggest more future defaults, and that loans of higher grades default less. It is surprising that income doesn't seem to have an effect on default, as you would assume higher earners would be able to pay back loans more than lower earners.

Centers for K-means(left) and Gaussian Mixture Modeling(right):



```
        PC1         PC2             [,1]        [,2]
1  1.077351 -0.09993375   [1,] -2.311974  0.20827427
2 -2.390720  0.22176032   [2,]  1.050781 -0.09465963
```

Discussion:

In my analysis the results for k-means and GMM on the clustering of the first two principal components were very similar. They both had error rates around 70%, correctly labeling around 10,000 observations correctly. The centers were also approximately the same around (1.1,-2.3) for PC1 and (-.09, .2) for PC2. The results of these clustering methods are similar because looking at the visual cluster plots, the clusters are spherical and k-means generally performs as well as or better than gaussian mixture modeling when clusters aren't elliptical. The findings of PCA showed some variables that contributed mostly to the variability in the first couple of principal components. These variables should have large differences in their means and counts in each of the clusters. It would be worth exploring the effect of the variables through a logistic regression on the principal components with reduced dimensionality in order to see how these predictors affect the exact probability of default. There is concern that I should have used three clusters instead of two, looking at either of the cluster visualizations there are three pretty distinct clusters, but my purpose of analysis was to find two meaningful clusters based on that status of a loan. That is to cluster the data into the two categories, then understand the distribution of each cluster to understand how the predictors contribute to each cluster. It may be interesting to explore an option with three clusters and examine the third cluster which may be a mix of defaulted and paid loans. However, the dataset was very imbalanced with only 20% of loans actually having defaulted, and this could have also contributed to the elbow method suggesting extra clusters.

<u>Analysis of Question 2:</u>

<u>Motivation:</u>

There are a number of factors for a company to consider amongst applicants about their demographic characteristics when deciding whether or not they can be approved for a loan. Understanding the dependence and independence between variables can better predict the probabilities of defaults. Having an understanding of the underlying joint probability distribution of all the factor variables can uncover insights on the probabilities of default given the occurrences of other factors.

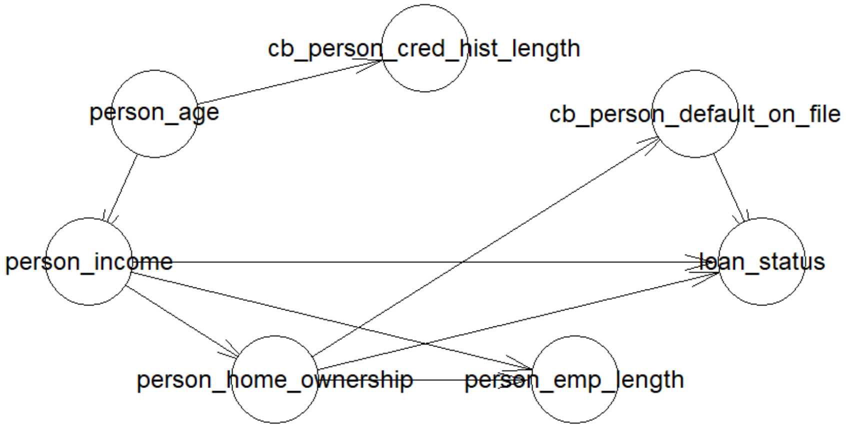
<u>Methodology:</u>

I created a graphical model more specifically, a Bayesian Network in order to map the relationships between the demographic variables and the *loan_status* variable. In order to first determine the structure of the Bayesian Network I ran a Tabu learning algorithm of the variables in discrete form. Then, I fit a Bayesian Network on the structure determined through Min-Max Hill Climbing. I used Min-Max Hill Climbing because the constraint-based approaches such as Incremental Association (iamb) produced a network that wasn't acyclic with undirected links, and the hybrid approach of min-max hill climbing produced an optimized network with highest score. In order to learn a Bayesian network with discrete components, I had to convert the continuous variables to discrete values. This was done through binary coding the variables to a value of "yes" or 1 if they contained a value above the median, and value of "no" or 0 for values less than the median of each respective continuous variable. In future studies or examinations I would be interested in exploring different ranges for the discrete coding or exploring gaussian components and learning parameters of their conditional joint distributions instead.

<u>Results:</u>

<u>Structure Learning for Bayesian Network</u>

The variables I selected for use in the Bayesian network relating to demographic/personal factors of the applicant are **age**, **income**, **home ownership**, **employment length**, **default** on **file**, **credit history length**, and **default status**. In order to learn the structure of the Bayesian Network, Min-Max Hill Climbing learning was used on the data frame, and the resulting structure was produced.

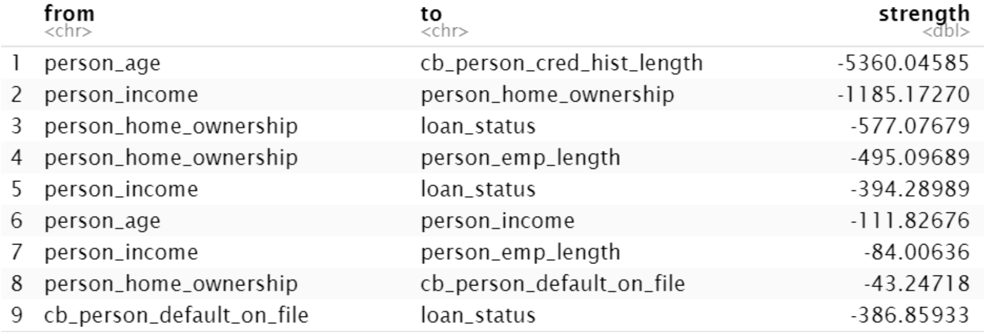


This structure contains 7 nodes, 9 arcs, is acyclic, and has no undirected arcs. Viewing the DAG (Directed Acyclic Graph) there are some conditional dependencies between parent and child nodes that make sense at first glance given background knowledge of loans and demographic statistics. Specifically credit history length having a parent node in age, income being a parent node of home ownership, income being a parent of loan status, and default on file a parent of loan status. Intuitively, most of these make sense because the older you are the more likely you are to have help a credit card or some form of credit for longer, higher incomes more likely to pay off loans or own a home, and if you have ever defaulted you are more likely to default in future or default on current loans.

Loan Status Markov Blanket

For the purposes of understanding how variables affect and relate to loan status, it is important to look at the markov blanket of the loan status node. The markov blanket consists of parents, children of the node, and the other parents of the children, and for loan status the markov blanket is only the parents of loan status, which are default on file, income, and homeownership. This gives us the information that age, employment length, and credit history length are conditionally independent of loan status, that is they are independent of loan status when they are conditioned on income, default file, and home ownership being known. This is important because it allows us to understand that the probability of a default/ the distribution of loan status isn't directly influenced by a person's age, credit history, or employment length, but only the states of income, home ownership, and default history are needed to evaluate the probability/ distribution.

Arc Strengths

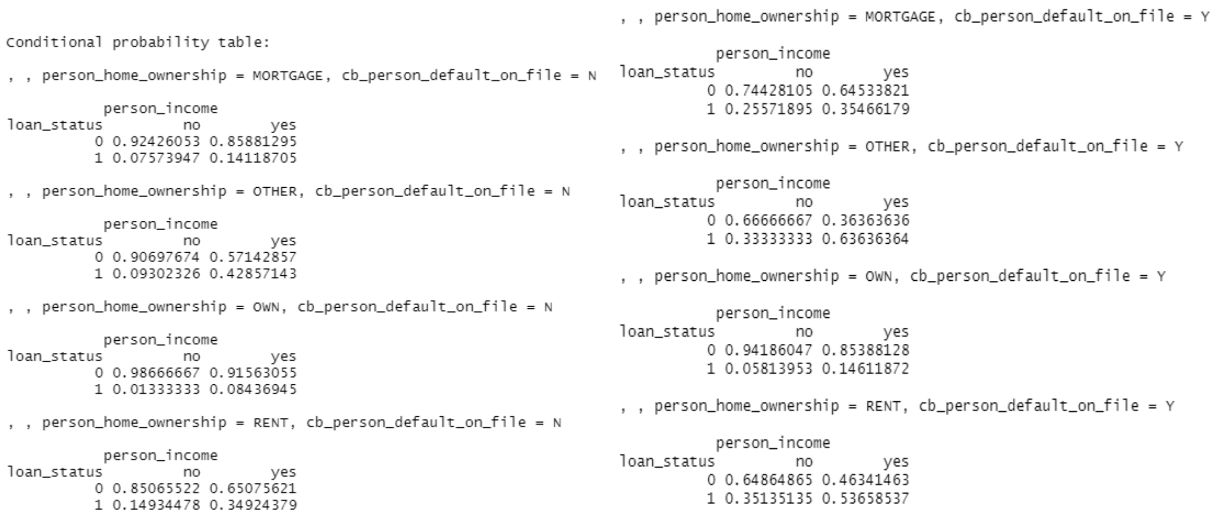


	from <chr>	to <chr>	strength <dbl>
1	person_age	cb_person_cred_hist_length	-5360.04585
2	person_income	person_home_ownership	-1185.17270
3	person_home_ownership	loan_status	-577.07679
4	person_home_ownership	person_emp_length	-495.09689
5	person_income	loan_status	-394.28989
6	person_age	person_income	-111.82676
7	person_income	person_emp_length	-84.00636
8	person_home_ownership	cb_person_default_on_file	-43.24718
9	cb_person_default_on_file	loan_status	-386.85933

According to the table the strength of dependency between the credit history and age is the strongest. The negative value represents a loss in the network score if the arc is removed from the network, so all arcs in the network contribute to a positive score increase.

Fitting Parameters for Conditional Probabilities:

Next, I fit the Parameters of the Bayesian Network based on the structure to obtain the conditional probability tables for each of the 7 nodes. The node of interest **loan status** has three parent nodes: default on file, income, and home ownership. The conditional probability table for loan status is below:



```
Conditional probability table:

, , person_home_ownership = MORTGAGE, cb_person_default_on_file = N

          person_income
loan_status         no        yes
          0 0.92426053 0.85881295
          1 0.07573947 0.14118705

, , person_home_ownership = OTHER, cb_person_default_on_file = N

          person_income
loan_status         no        yes
          0 0.90697674 0.57142857
          1 0.09302326 0.42857143

, , person_home_ownership = OWN, cb_person_default_on_file = N

          person_income
loan_status         no        yes
          0 0.98666667 0.91563055
          1 0.01333333 0.08436945

, , person_home_ownership = RENT, cb_person_default_on_file = N

          person_income
loan_status         no        yes
          0 0.85065522 0.65075621
          1 0.14934478 0.34924379

, , person_home_ownership = MORTGAGE, cb_person_default_on_file = Y

          person_income
loan_status         no        yes
          0 0.74428105 0.64533821
          1 0.25571895 0.35466179

, , person_home_ownership = OTHER, cb_person_default_on_file = Y

          person_income
loan_status         no        yes
          0 0.66666667 0.36363636
          1 0.33333333 0.63636364

, , person_home_ownership = OWN, cb_person_default_on_file = Y

          person_income
loan_status         no        yes
          0 0.94186047 0.85388128
          1 0.05813953 0.14611872

, , person_home_ownership = RENT, cb_person_default_on_file = Y

          person_income
loan_status         no        yes
          0 0.64864865 0.46341463
          1 0.35135135 0.53658537
```

The table is broken into the different combinations of the levels of each variable. From this table the probability of default conditioned on the three parents is listed for each value of *loan_status*. The value of "0" for *loan_status* indicates a non-defaulted loan, and the value for "1" indicates a defaulted loan. The value "yes" for *person_income* indicates the applicant's income was larger than the median, and the value "no" indicates a value of income that is below the median. For a value of "Y" for *cb_person_default_on_file* or yes for a default on file the loan status probability significantly increases. These values for the conditional probabilities all align with what intuition would expect given previous knowledge of loan data and demographic data.

Discussion:

The resulting Bayesian network and conditional probability tables found from learning a network on the data allows us to query the Network in order to determine the probability of events on the network given different amounts of known evidence. For the purpose of our study It may be interesting to predict the probability of a default using different values for each of the parameter nodes. For example, the inferred conditional probability of a person with a default on file, higher than median income, and Mortgage is .3538. The conditional probability of a person using the evidence of just renting a home and a previous default on file is .4557. Further analysis of this data using Bayesian Networks should explore using Gaussians instead of discrete coding the variables, or at least code the discrete variables to more than two levels to allow for more specific predictions of the conditional probabilities. The strengths of the relationships could also be explored further because the parent nodes of loan status didn't necessarily have the largest negative strength, it may be interesting to look at correlations between these variables or hypothesis testing with them in a regression model to explore their significance with loan status as a response.

Appendix

Sources:

- https://www.r-bloggers.com/2015/02/bayesian-network-in-r-introduction/#:~:text=Bayesian%20networks%20(BNs)%20are%20a,other%20unsupervised%20machine%20learning%20techniques.
- https://kuanjh123.medium.com/markov-blanket-8aaa416495c3
- https://www.bnlearn.com/documentation/man/structure.learning.html
- https://www.bnlearn.com/examples/fit/
- https://www.bnlearn.com/about/thesis/thesis.pdf
- https://www.bnlearn.com/about/teaching/slides-bnshort.pdf
- https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/
- https://medium.com/@zullinira23/implementation-of-principal-component-analysis-pca-on-k-means-clustering-in-r-794f03ec15f
- https://towardsdatascience.com/clustering-out-of-the-black-box-5e8285220717
- https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data

Code:

```
#QUESTION 1
        library(tidyverse)
        d<-read.csv("C:\\Users\\User\\Downloads\\credit_risk_dataset.csv")
        summary(d)

        d2<-na.omit(d)
        summary(d2)

        d3<-d2[-which(d2$person_age>=90 | d2$person_emp_length > d2$person_age-15),]
        d3<-d3[-which(d3$loan_percent_income==0),]
        summary(d3)
        d_clean <- d3 %>% mutate_if(is.character, as.factor)

        #Test/Train Split
        library(dummies)
        d_dummy<-dummy.data.frame(d_clean,names=c("person_home_ownership","loan_intent","loan_grade","cb_person_default_on_file"))
        str(d_dummy)

        set.seed(100)
        sample.data<-sample.int(nrow(d_dummy), floor(.5*nrow(d_dummy)), replace = F)
        train<-d_dummy[sample.data, ]
        test<-d_dummy[-sample.data, ]
```

```
#PCA to reduce dimensions for Clustering
train.x<-train[,-23]
train.y<-train[,23]
prin_cmp<-prcomp(train.x,scale.=TRUE)



#loadings
prin_cmp$rotation[,1:2]

library(factoextra)
library(ggfortify)
autoplot(prin_cmp,label=FALSE,data=train.x,loadings=TRUE,loadings.label=TRUE,loadings.label.size=3,shape=FALSE)

fviz_eig(prin_cmp,addlabels=TRUE)

biplot(prin_cmp,scale=0)

stdev<-prin_cmp$sdev
var<-stdev^2

prop_varex<-var/sum(var)

plot(prop_varex,type="b")
plot(cumsum(prop_varex),type="b",ylab="Proportion of Variance",xlab="Number of Principal Components");abline(h=.8,col="red")
#around 20 principal components


#elbow method
set.seed(20)
wss<-NULL
for(i in 1:8){
  fit<-kmeans(prin_cmp$x[,1:2],centers=i)
  wss<-c(wss,fit$tot.withinss)
}
plot(1:8,wss,type="o")

#kmeans clustering


fit <- kmeans(prin_cmp$x[,1:2],2,nstart=25)
```

```
fit$centers

#visualization of clusters using first two principal componenets.
fviz_cluster(fit,prin_cmp$x[,1:2])

#Comparing the actual cluster labels to the training kmean labels
act<-train.y+1
clustlabs<-fit$cluster
df1 <- data.frame(train.x,actual=act,clust_lab=clustlabs)

sum<-0
for (i in 1:nrow(df1)){
  if (df1$actual[i] == df1$clust_lab[i]){
    sum<-sum+1
  }
  else{
    sum<-sum+0
  }
}
#success rate
k_correct_rate<-sum/nrow(df1)
#error rate


#summaries of variables
cluster_summaries<-df1%>%group_by(clust_lab)%>%summarise(ratemean=mean(loan_int_rate),default_count=sum(cb_person_default_on_fileY),mean_inc=mean(person_income),mean_age=mean(person_age),a=sum(loan_gradeA),b=sum(loan_gradeB),c=sum(loan_gradeC),rent=sum(person_home_ownershipRENT),mortgage=sum(person_home_ownershipMORTGAGE));cluster_summaries

#Gaussian Mixture Modeling
library(mclust)
library(ClusterR)

#normalize data
gm_train2<-scale(train.x)

#gaussian mixture model
gm_model2<-GMM(prin_cmp$x[,1:2],gaussian_comps=2)
```

```
gclusterlabels<-predict(gm_model2,prin_cmp$x[,1:2])

for(i in 1:length(gclusterlabels)){
  if(gclusterlabels[i]==1){
    gclusterlabels[i]<-2
  }
  else{
    gclusterlabels[i]<-1
  }
}

df2<-data.frame(actual=act,glab=gclusterlabels,prin_cmp$x[,1:2],train.x)

#centers
gm_model2$centroids

sum1<-0
for (i in 1:nrow(df1)){
  if (df2$actual[i] == df2$glab[i]){
    sum1<-sum1+1
  }
  else{
    sum1<-sum1+0
  }
}

library(ggplot2)
ggplot(data=df2,aes(x=PC1,y=PC2,group=gclusterlabels,color=gclusterlabels))+geom_point()+labs(title="Cluster Plot GMM")

cluster_summaries<-df2%>%group_by(glab)%>%summarise(ratemean=mean(loan_int_rate),default_count=sum(cb_person_default_on_fileY),mean_inc=mean(person_income),mean_age=mean(person_age),a=sum(loan_gradeA),b=sum(loan_gradeB),c=sum(loan_gradeC),rent=sum(person_home_ownershipRENT),mortgage=sum(person_home_ownershipMORTGAGE));cluster_summaries
```

```
#QUESTION 2
#data
library(tidyverse)
d<-read.csv("C:\\Users\\User\\Downloads\\credit_risk_dataset.csv")
summary(d)
d2<-na.omit(d)
```

```
#converting to discrete values
med<-c()
for (i in 1:ncol(d2)){
  if(class(d2[,i])!="character"){
    med[i]<-median(d2[,i])
  }
  else{
    med[i]<-0
  }
}
for(i in 1:ncol(d2)){
  if(med[i]!=0){
    d2[which(d2[,i]< med[i]),i]<-"yes"
    for(f in 1:length(d2[,i])){
      if(d2[f,i]!="yes"){
        d2[f,i]<-"no"
      }
    }
  }

}

library(bnlearn)
d_clean <- d2 %>% mutate_if(is.character, as.factor)
d_clean$loan_status<-as.factor(d_clean$loan_status)

d_clean2<-d_clean[,c(1,2,3,4,9,11,12)]


#hill climbing
try2<-hc(d_clean2)

#min max hill climbing
try5<-mmhc(d_clean2)

#visualizing structures
plot(try2)
plot(try5)

#structure information
try5

#fitting distribution for nodes using maximum likelihood estimation
fitbn<-bn.fit(try5,data=d_clean2,method="mle")
```

```
fitbn$loan_status

fitbn

#predicting values/events given values of certain nodes
cpquery(fitbn,(loan_status=="1"),(cb_person_default_on_file=="Y" &
person_home_ownership=="RENT"))

library(Rgraphviz)
library(bnlearn)
strength.plot(try5,arc.strength(try5,d_clean2))
```