EMIS 7357

Analytics for Decision Support

Project # 2

Ilumexico

By:

APOORVA JAIN

Executive summary

Ilumexico have been collecting basic data for each one of our customers in other to understand better the market we serve. Since 2017, we have included a set of new data points that gives us the opportunity to better segment the base of the pyramid market. With that data, we can begin to understand how each of this data points influence the monthly instalment capacity, payment behavior and overall loan performance. This analysis will help them to refine there financing options so they can adequate better to the economic capabilities, cash flow and opportunities; and improve our overall portfolio performance. Well after analyzing I can conclude that our model for prediction is almost 62 % accurate but, on the training, sets its much higher that is 68% that usually happens. We have analysed various other factor with the help of cluster analysis, logistic regression and used TABLEAU software. I can also comment on the pattern of our clients, rather talk about there background and there payments and some sociodemographic relation.

Table of contents

Contents

List of Figures

Introduction

Introducing Ilumexico company before we start analyzing the different aspects to achieve the goals. It's a Mexican social enterprise that was funded in 2009 by few engineering students. They had a goal of minimizing the energy shortage with the help of solar system. They designed it according to the different aspects of rural life that are as follows

1. Family
2. Economic activities
3. Public spaces

According to that the business model of ilumexico consist :

1. Product configuration for off grid needs
2. Last mile customer care
3. Inclusive financing
4. Exceptional customer experience

They have 80 employees which provides 14600 households and energy t more than 62000 people distributed in more than 1350 rural communities. We already had two rounds of financial investments that took place in 2012 and 2017 with international social impact funds. Our company is also BCorp and GIIRS certified enterprise.

We also have flexible financing products that are as follows:

1. Cash payment
2. Pure credit

3. Service monthly instalment
4. Pay to own technology

We innovated the above flexibility so that we can provide multiple choices for every context and every economic circumstance.

Analysis

1. Relation between sociodemographic data and payment behavior

Relation between sociodemographic data and payment behavior that gives a relation between what

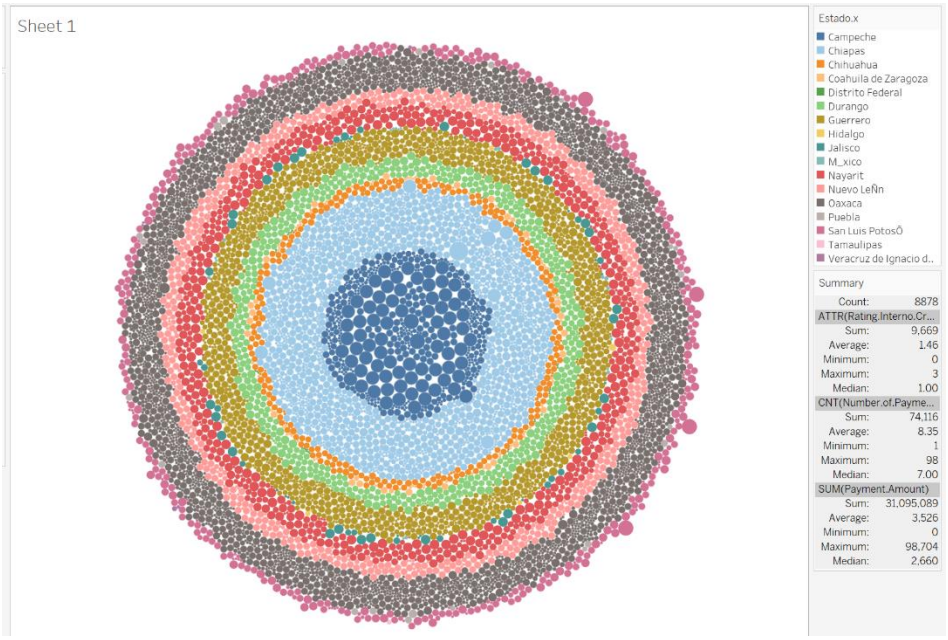


Figure 1 RELATION SOCIODEMOGRAPHIC AND PAYMENT BEHAVIOR

should be the internal credit score. We evaluated this Tableau software and visualization is provided in the figures as shown generally and with respective to the internal credit rating . The different colors tell me the states which can be seen and interpreted by the legends in the figure. There are approximately 2500 customers with credit score of 3 , 744 with the credit rating of 2 almost similar credit rating1 but for credit rating of 0 there are almost 2700 customers. There are

few customers they still don't have the credit rating and hence there rating is null . shown below in the figure.

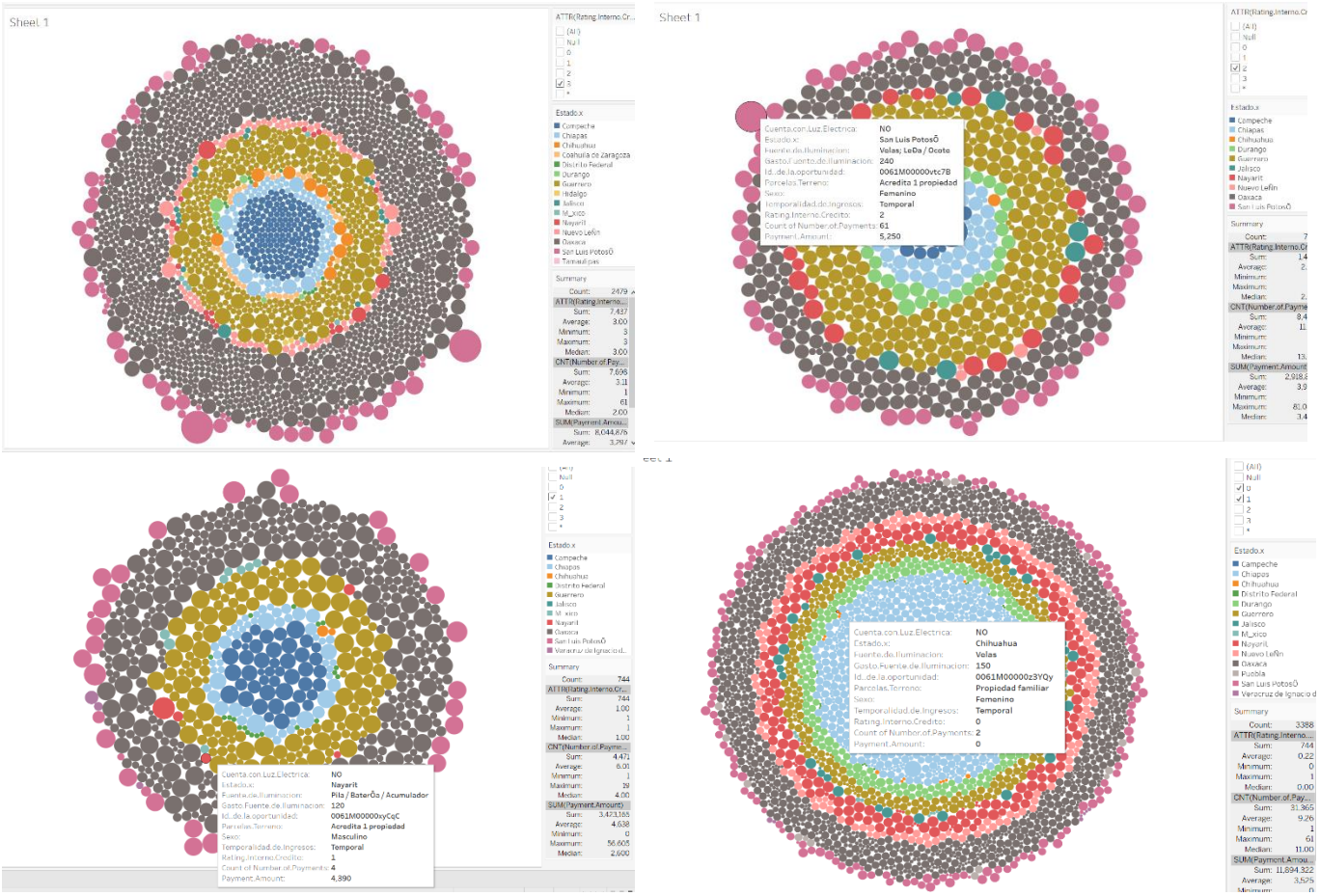


Figure 2 CREDIT RATING 3,2,1,0

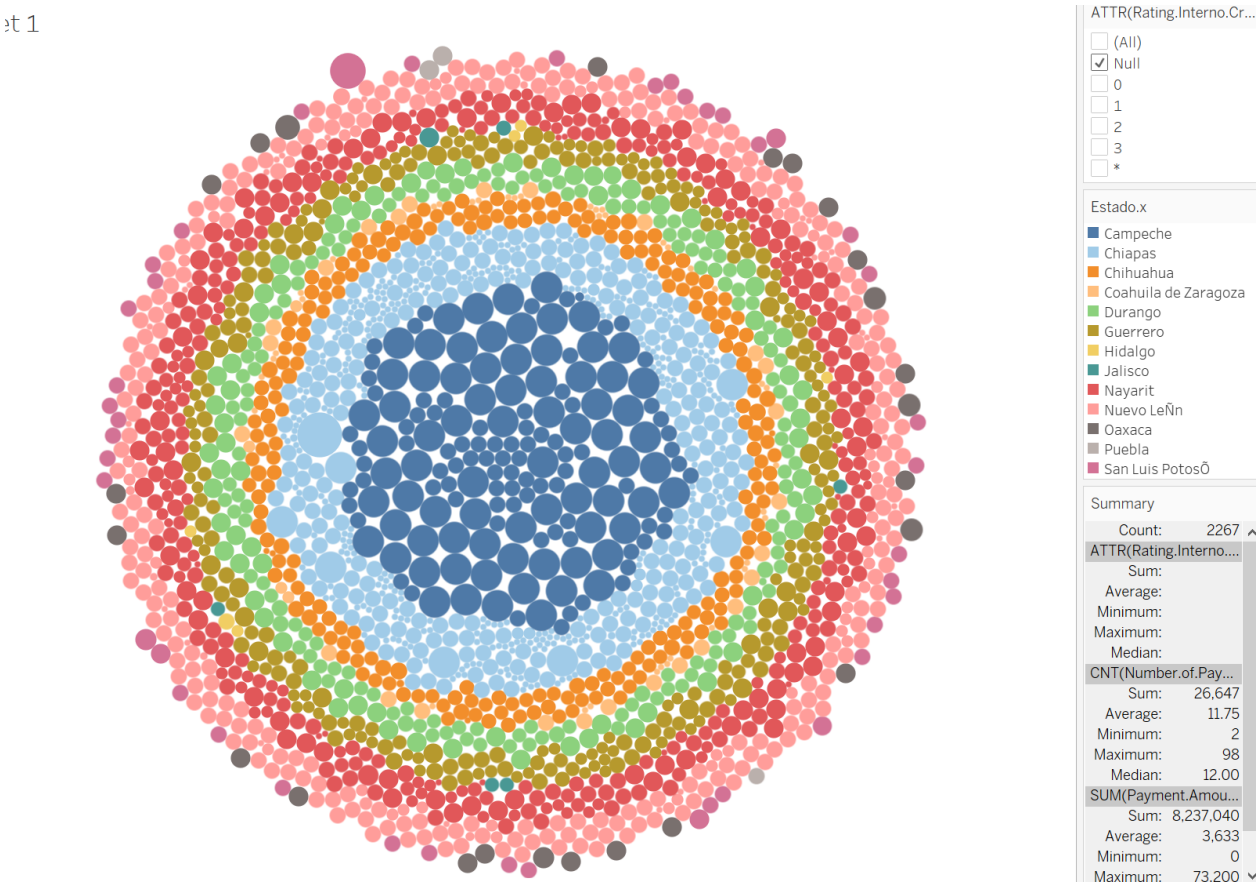


Figure 3 Null credit rating (didn't received)

Just using our mouse and moving around the visulaization we can know the number of defaulters according to the sociodemographic data and many more things accoridng to our reuirements which allows us to know where do we have to expand our business and where do we lack.

2.Create logistic regression for valid credit score or payment

We have used logistic regression and build a model to predict and validate payment or the payment probability which is shown below in the figure .We have used Paid as the output should be 0 or 1 (binary variable) to use the logistic regression .We have to refine the model many a times to get this significant model which shows total balance , paid balance and credit term months are statistically significant though our internal credit rating is also significant but not as much as the above factors are. So this model do validate the credit score or the payment as rating will only be good when the balance is paid and the credit term is more and that totally signifies that total balance, this all together makes this model valid.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              4.032e-01  1.552e-01   2.598  0.00939 **
SexoFemenino            -1.501e-01  6.043e-02  -2.485  0.01297 *
SexoMasculino           -1.441e-01  5.208e-02  -2.767  0.00566 **
total.balance           -3.888e-04  2.099e-05 -18.527  < 2e-16 ***
paid.balance             1.549e-04  1.049e-05  14.774  < 2e-16 ***
credit.term.months      -4.794e-02  2.924e-03 -16.398  < 2e-16 ***
months.passed           -8.652e-03  4.304e-03  -2.010  0.04442 *
Rating.Interno.Credito -3.585e-02  1.765e-02  -2.031  0.04228 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4 Logistic regression model

From the training model that we build on the data we collected basically past data so that we can evaluate our model we calculated our False positive rate and true positive rate with the help of these rates we analyse are accuracy.

	FALSE	TRUE
0	5201	6004
1	1556	6739

Accuracy for this model is equivalent to 61% which is not that good . Prediction curve for training sets is

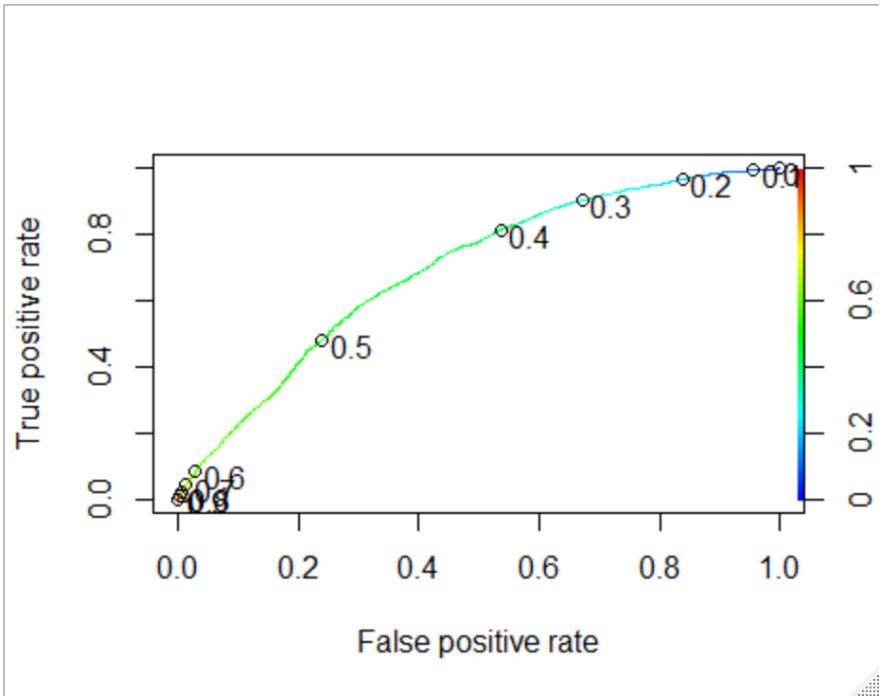


Figure 5 AUC curve training data sets

shown in figure below.

The AUC Value : 0.691 that is not bad but its realistic.

Note: threshold value t is considered to be 0.4

3.Relevant prediction or evaluation

To evaluate or to predict our model we created two different sets mainly training and testing sets but for future we generally test our model on testing sets . We also have seen what happens on our training model previously.

Testing model related to internal credit rating is highly significant as we can see from the model shown below.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             6.355e-01  2.115e-01   3.004  0.00266 **
SexoFemenino           -1.066e-01  8.142e-02  -1.309  0.19045
SexoMasculino          -3.031e-01  7.175e-02  -4.224 2.40e-05 ***
total.balance          -3.864e-04  2.877e-05 -13.432  < 2e-16 ***
paid.balance            1.678e-04  1.421e-05  11.809  < 2e-16 ***
credit.term.months     -5.751e-02  4.048e-03 -14.208  < 2e-16 ***
months.passed          -1.129e-02  5.913e-03  -1.910  0.05614 .
Rating.Interno.Credito -9.603e-02  2.414e-02  -3.978 6.96e-05 ***
---
```

Figure 6 Testing model for prediction

For prediction purposes we need to calculate the accuracy to validate our model for further evaluation.

	FALSE	TRUE
0	2782	3251
1	821	3646

AUC Value : 0.688 which good not bad as AUC value for an accurate model is 1. Here is the AUC curve shown below.

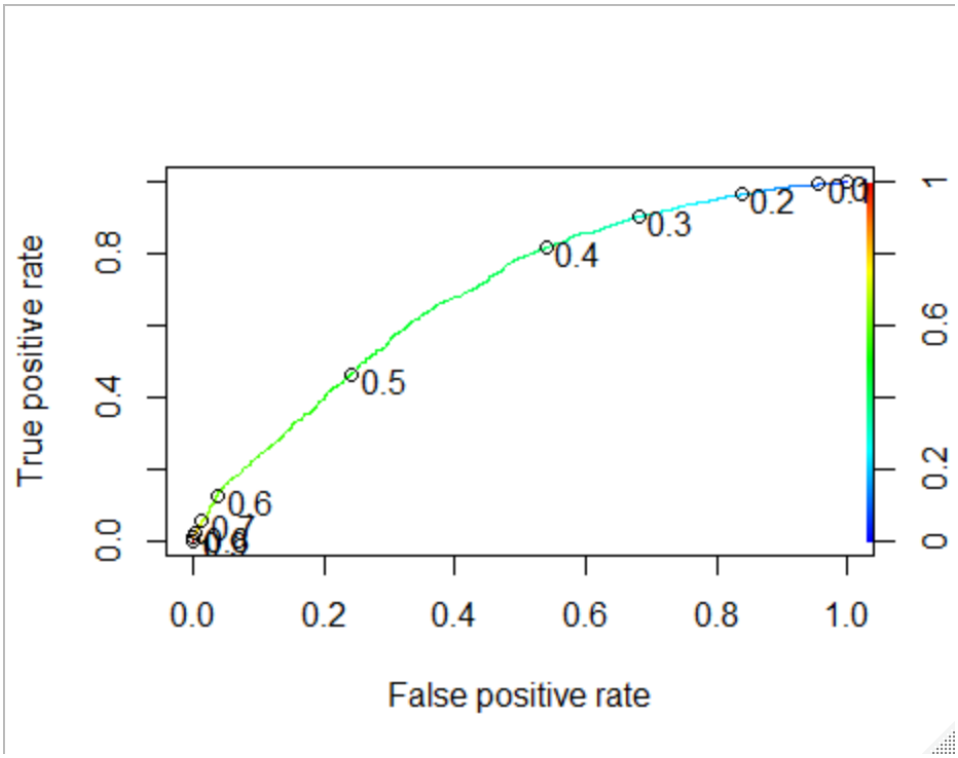


Figure 7 AUC curve on testing dataset for prediction

4.Credit rating reliable or not

The rating are reliable are not we can only predict by the help of our testing data sets that we used above to predict with the help of AUC curve and TPR , FPR. For the prediction we used our testing set and accuracy on that is calculated as shown below and compared to its baseline accuracy.

Accuracy for our model is (2982+3946)/(3251+821+2782+3646) =6428/10500=0.666 that turns out to be 66% though its not efficient but cannot say its bad for prediction purposes. Further we can be reliable but not totally reliable on these credit ratings.

5.Correlation between customer compliance and payment schedule

We analyzed the correlation between the customer and payment schedule by the data sets which we tried to explore it according to our motives. I referred few articles about the customer compliance how they deal with it , there are few factors that I took into consideration to see the correlation between the two in our company that are : history of client that can be reflected by the situation of the client is he/she in debt or not , do they have late payment or how much balance is due , are they defaulters in any case and the best factor that I took into consideration is internal credit rating of our system. The output of my correlation can be seen in the figure below.

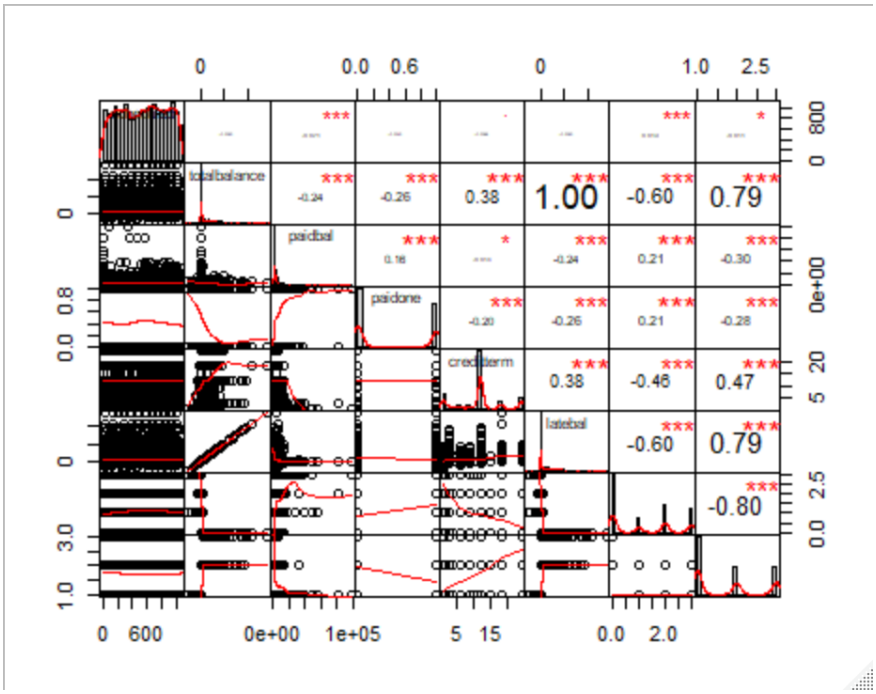


Figure 8 Correlation graph

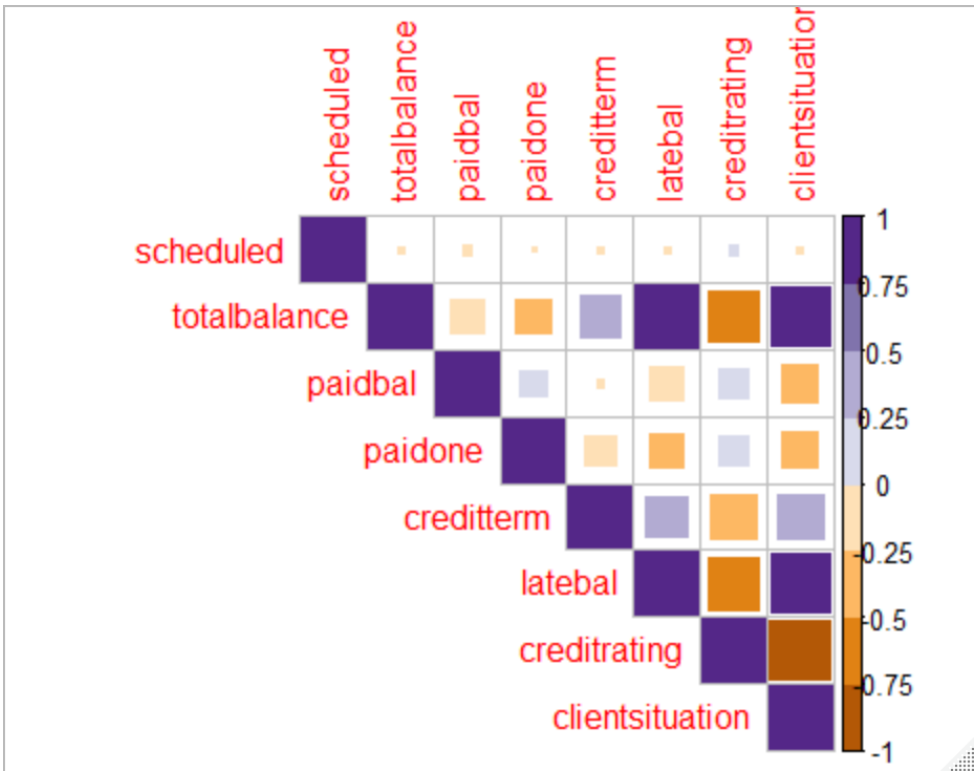
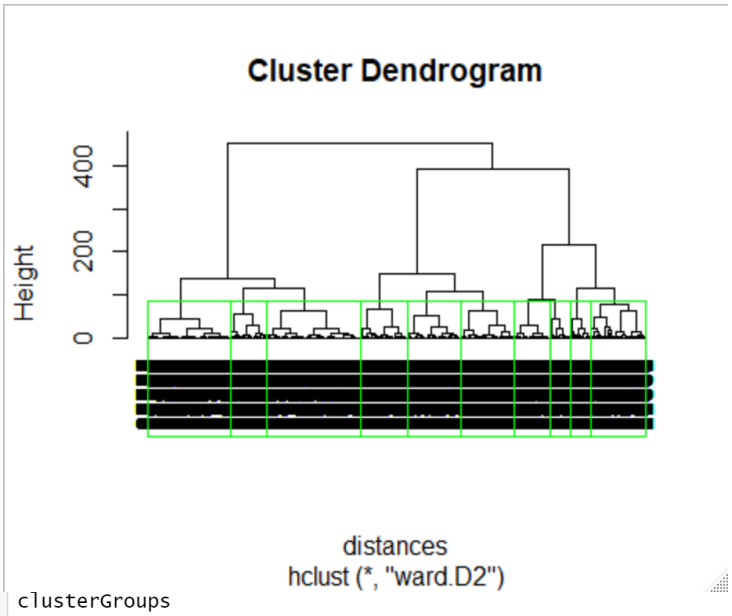


Figure 9 Better visualization of correlation

So from here we can say that internal crediting rating is the best customer compliance for us to judge there behavior though there are various other factor having the high correlation factor client situation and total balance, situation of client dealing with debt or late balance these all factors effect the scheduling. All these factors indirectly or directly effect the schedule.

6.Cluster Analysis

So here I analyzed the clients with the help of cluster analysis. I preferred using hierarchical clustering taking the number of clusters equivalent to 10 into consideration also I took some factors into the consideration for the cluster analysis from the data provided so that I can predict what client do, which type of client is interested they are from which state , locality , type of there income ,age and the main thing there credit rating in our systems. Here I provide a figure of dendogram of clusters in figure shown below with the distributed data in each cluster.




```
clusterGroups
   1    2    3    4    5    6    7    8    9   10
3184 4952 3369 3263 2800 2223 5634 2166 1209 1200
> |
```

Figure 10 Dendogram of clusters of client and grouping of clusters.

Conclusion

Well after analyzing I can conclude that our model for prediction is almost 66 % accurate but on the training sets its much higher that is 68% that usually happens. We also have analyzed various other factor with the help of cluster analysis, logistic regression and used TABLEAU software. I can also comment on the pattern of our clients, rather talk about their background payments and some sociodemographic relation. The model is reliable so we can comment the internal rating is reliable and we can use it in future for evaluating till than all the clients who still don't have their ratings will also be included so that we might know how other factors are going to be impacted.

Reference

1. http://www.cfo.com/credit-capital/2013/01/when-your-big-customer-wants-to-pay-late/
2. https://datascienceplus.com/hierarchical-clustering-in-r/
3. https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/