Business Analytics Project 2020

Credit Worthiness

NAME: Mawali Knight

DATE: 6th September, 2020

LECTURER: Mr. Harris

COURSE: PDLL115-Business Analytics

INSTITUTION: Centre for Professional Development and Lifelong Learning (University of the West Indies, Cave Hill Campus)

Table of Contents

Abstract

This paper explores a data set by looking at the factors that promote credit worthiness and credit unworthiness by analyzing related and unrelated variables about people. In addition, a support vector machine will be utilized by using its logistic regression feature. This feature is used to produce a model that can be used to predict credit worthiness, herein falling into the category of predictive analytics.

Introduction

Recently, credit risk assessment has been trending across the globe. Wang et al (2011) believe this increased attention has been mainly due to the weakness of current risk management methods that have been shown from the 2008 financial crisis and the increasing demand for consumer credit. As a result, there has been an increasing importance of credit scoring as financial institutions now shy away from the traditional methods (Huang, Chen and Wang, 2007). This paper attempts to develop a credit scoring model using the tools of a support vector machine, especially through the use of its logistic regression feature.

Literature Review

According to Zhou et al (2009), credit scoring is a way of assessing the credit risk of loan applicants with their corresponding credit score which is acquired from a credit scoring model. It can also be defined as the statistical method undertaken to convert the data into rules that can be utilized for the guiding of credit granting decisions (Eisenbeis, 1978). A credit score is a value that can represent whether an applicant is creditworthy or not and it is based on the evaluation of an applicant's characteristics (Zhou et al, 2009). Since there has been a swift growth in the credit industry, credit scoring models have been significantly utilized for evaluating credit admissions (Thomas, 2000).

The support vector machine (SVM) developed by Vapnik (2000), is a successful classification technique used for credit scoring (Lee, 2007).

Methodology

The Credit Risk dataset was obtained from a local lending agency and it consists of 425 records and 27 attributes. It classifies individuals based on if they are creditworthy or not.

Firstly, the Credit Risk dataset will be loaded and then will be split randomly into a test set and training set. The first SVM model will be trained is a linear SVM model. In addition, the test percentage will be changed to 10% and consequently the training set will carry the remaining 90% of the data. In addition, Lambda will be set to 3 as this was found to improve accuracy based on personal experimentation. Finally, the performance of the model on the training and test dataset will be gauged. The SVM algorithm was carried out in the Octave program.

Empirical Results and Analysis

It is well known that credit scoring is a daunting task as credit data is not easily separable. This is partly due to the fact that credit data tends to neglect the use of qualitative data.

Summary Statistics

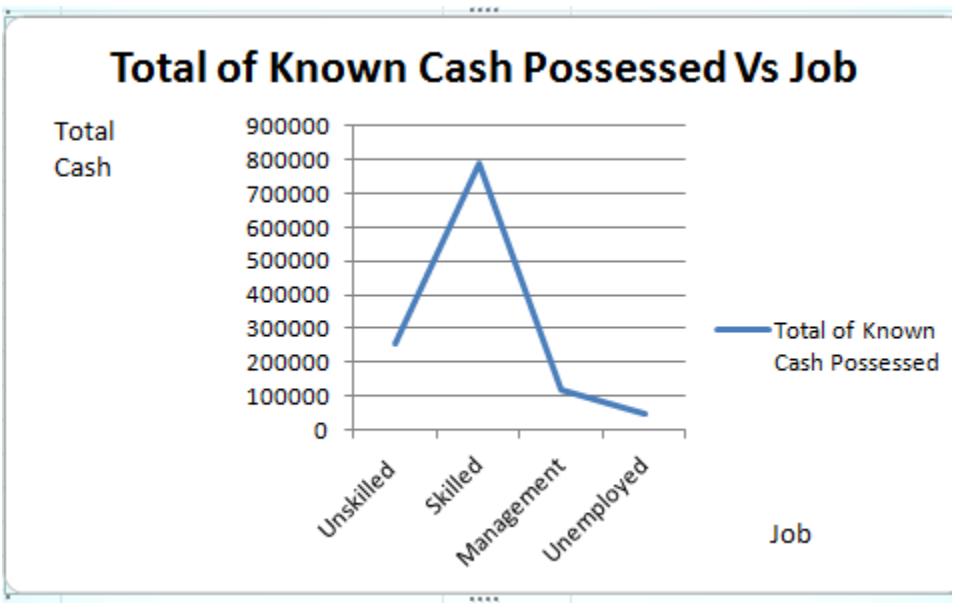
- The majority of individuals are male
- Most persons are single
- The mean age is 34 (Rounded to the nearest whole number)
- The majority of persons own where they reside and also most persons are skilled in terms of their labor classification
- The means for the monetary balances are: $1,048.01 (Checking balance), $1,812.56 (Savings balance), $2860.58 (Total of known cash possessed (Checking plus Savings))
- The modal the monetary balances are: $0.00 (Checking balance), $0.00 (Savings balance), $0.00 (Total of known cash possessed (Checking plus Savings))
- The median monetary balances are: $0.00 (Checking balance), $536.00 (Savings balance), $836.00 (Total of known cash possessed (Checking plus Savings))

Correlations

- Both Savings and checking balances are negatively related to credit risk (This would suggest that these do not contribute much to a high credit risk). This is unexpected since cash normally is indicator of wealth.
- Months that people were customers is positively related to credit risk (This would indicate that months the customers have been members plays a role in increasing credit risk). This is unexpected as this would suggest new members tend to be more creditworthy than established members.
- Months employed appear to be negatively related to credit risk. This is expected as the longer people are employed it is expected that will accumulate a lot of liquity and current assets.
- Gender is negatively related to credit risk. This would seem to be saying that men are better in terms of credit risk than women.
- Carrying the relationship status of single is negatively correlated to credit risk. Therefore, this would suggest that people who are single appear to be poor in terms of credit risk). This is unexpected as these people normally only have one source of income and they do not have the leverage that comes with an intimate union.
- People who carry the relationship status of divorced are positively correlated to credit risk (Carrying the title of divorced appears to increase credit risk).
- Being skilled or being in a management position both carry positive relationships to credit risk. Hence, this seems to indicate that people in these positions are generally not creditworthy)

- The number of years a person resides at a residence has a positive relationship to credit risk. This is unexpected as a fixed and established place of abode is generally a sign of stability.
- The relationship of being a renter is positively related to credit risk. This is expected as these people are missing the collateral of an owned residence.

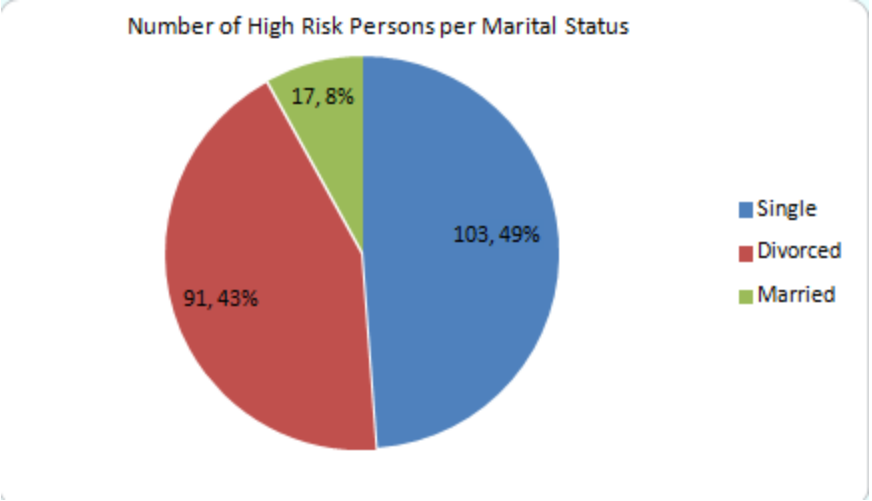
Visual Statistics



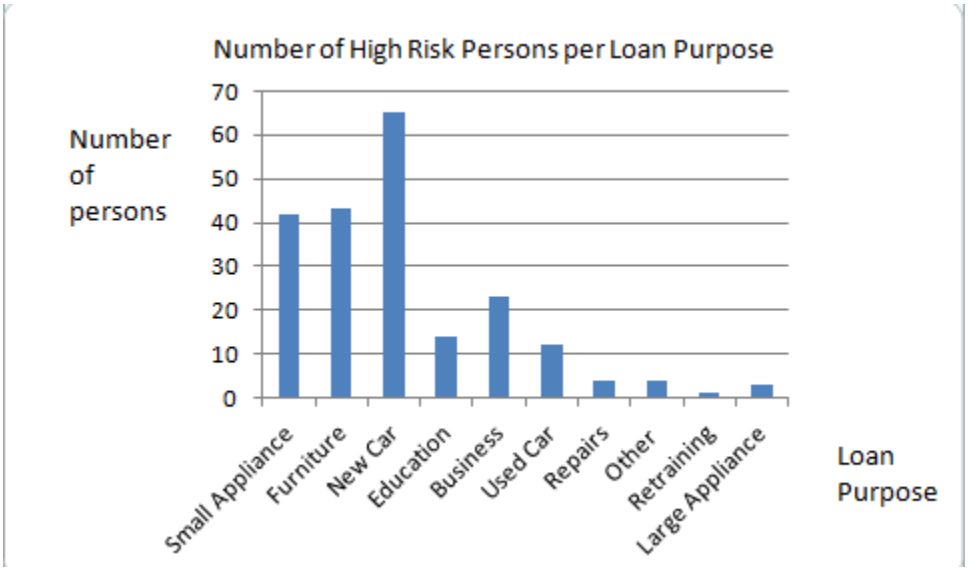
This line graph shows us that skilled workers tend to possess the lion share of the total known cash possessed (Checking plus savings). They are subsequently followed by unskilled employees, management employees and then unemployed persons.

Number of High Risk Persons	Total Number of Persons
211	425

This table shows that out of 425 persons 211 people (49.65%) are not creditworthy.



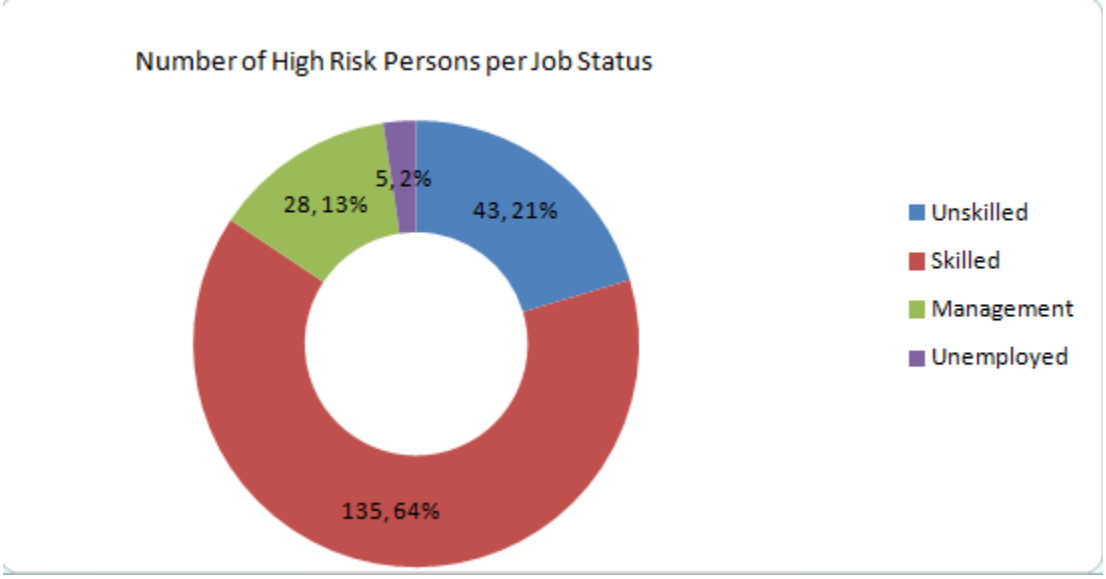
This pie chart shows that single people are the riskiest people to lend to. They are closely followed by divorced persons and then married people.



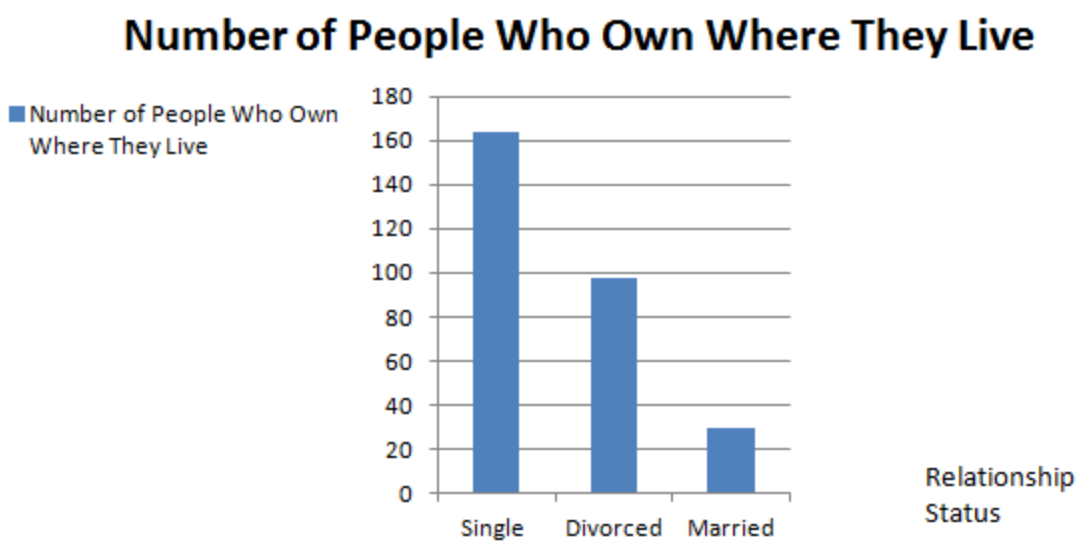
This bar chart shows that the riskiest loans to lend are ones for new cars. After this, in terms of riskiness, comes loans for furniture, small appliances, business, education, used cars, repairs, miscellaneous purposes, large appliances and retraining.

Number of High Risk Persons Based on Living Sitution	
Own	Rent
131	49

This table unexpectedly shows that homeowners have a higher proclivity with regards to being riskier to lend to as opposed to renters.



This doughnut chart highlights that skilled labourers are the riskiest to lend to. This is contradictory as this category also possesses the great accumulation of liquid cash.



This bar chart defies what is seen on a day to day basis in most countries. Unexpectedly, single persons from the sample own most of the homes, followed by divorced people and then comes married persons.

Number of People Over or Equal to 34 Who Are High Risk	Number of People Under 34 Who Are High Risk
82	129

As it is known that the mean age of the sample is near 34, an assessment was undertaken to see if either people below or above this age are safer to lend to. Based on the findings, it can be concluded that people over are equal to 34 are generally safer to lend to as expected as wealth tends to be characteristic held by older people.

```
Train Accuracy: 68.062827

Program paused. Press enter to continue.
Test Accuracy: 69.767442
```

The model predicts accurately on the training dataset if a person is credit worthy 68.06% of the time based on this finding produced via the SVM.

The model predicts accurately on the test dataset if a person is credit worthy 69.77% of the time based on this finding produced via the SVM.

```
>> theta
theta =

  -0.0027627
  -0.3553264
  -0.2638717
   0.0926466
  -0.0370379
  -0.0819799
  -0.4681390
  -0.2820961
  -0.0119697
   0.1081282
  -0.0300370
  -0.1746033
  -0.1462121
   0.5064491
  -0.6089066
  -0.0513063
  -0.1214039
   0.1918713
  -0.1239729
   0.2015064
  -0.1367308
   0.0199548
   0.0254422
   0.3382552
   0.3452377
   0.3240379
   0.2786012
```

These are the values which make up the linear equation which predicts the creditworthiness of the sample.

e.g) Y=Bo+B1X1+B2X2+…

Actual) Y=-0.0027627+(-0.3553264 x X1)+(-0.2638717xX2)+…

Conclusion

Based on the findings the most ideal/ safest person to lend to looks like someone who is married, wants a loan for retraining purposes, is a renter, is unemployed and are over the age of 34. In addition, some undesirable characteristics based on the correlation report which should be taken into consideration are long customer patronage, divorcees, a skilled worker or a person in a management position, a person who has been residing where they live for a long period of time, and a renter. The feature of renting is contradicting but based on general knowledge and what is seen in everyday life renters are not as stable as homeowners. Furthermore, according to an article from the New York Times, titled Homeownership Equals Stability, homeownership is a generally a great indicator of a solid financial standing.

Acknowledgement

"I come as one but I stand as ten thousand". This statement from Maya Angelou could not hold truer as this dissertation took many moulding hands to produce. From my family and friends motivating me through the many long nights and days to me being ably and unwaveringly guided my lecturer Mr. Harris, I had reliable support systems.

With that being said and in successfully completing this project, I would first like to thank God for bestowing me with wisdom to present intellectual findings and sophisticated analyses.

Secondly, the steadfast support and time expended by Mr. Harris, my teacher, must and will not be forgotten, as he provided me with sound and helpful advice and guidance. Moreover, Mr. Harris tapped into and refined the potential I have as a business analyst from the time of inception of the research idea to the very end. As a result, complete, accurate and thorough research was produced to aid lending institutions. A big thank you goes out to Mr. Harris for all of his assistance.

Additionally, thanks goes out to my family and friends as family and friend encouragement played a big role in helping me to get across the finish line.

Lastly, I would like to thank all of the people who contributed to the success of my Business Analytics Project.

References

Eisenbeis, R, A. (1978) 'Problems in applying discriminant analysis in credit scoring models', *Journal of Banking & Finance*, 2(3), pp 205-219.

Huang, C, L., Chen, M. C., & Wang , C. J. (2007) 'Credit scoring with a data mining approach based on support vector machines', *Expert Systems with Apllications*, 33(44), pp 847-856.

Lee, Y. C. (2007) 'Application of support vector machines to corporate credit rating predicition'. *Expert Systems with Applications*, 33(1), pp 67-74.

Mayer, C. (16th July, 2014). *Homeownership Equals Stability*. The New York Times Newspaper

Thomas, L.C. (2000) 'A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers', *International Journal of Forecasting*, 16(2), pp 149-172.

Wang, Z., Yan, S. C., &Zhang, C. S. (2011) 'Active learning with adaptive regularization, *Pattern Recognition*, 44(10-11), pp 2375-2383.

Zhou, Ligang & Lai, Kin Keung & Yu, Lean. (2009) 'Credit scoring using support vector machines with direct search for parameters select ion', *Soft Comput*, 13(2), pp 149-155.