December 8, 2024

Elliot K. Hlorgbe

M00363771

Term Project

CIS 643-22-FA24 - Advanced business analytics

**Introduction**

In this project our objective was to analyze data provided by Dream Housing Finance, a company that provides home loans to customers across urban, semi-urban and rural areas. The data is based on the information a customer provides when completing an online application, the company wants to automate the loan qualifying process in real time. This project consists of multiple analyses designed to derive some understanding and patterns in the data. For this research we conduct descriptive and frequency analyses, correlation analysis, a cross table with chi-square test, comparison test, multiple regression analysis, and logistic regression analysis. The data was obtained from Kaggle.com by a “exploratory data analysis” keyword search which yielded the “Home loan approval” project uploaded by Prepinsta Technologies. The data consists of 13 fields made up of 8 categorical variables and 5 numerical variables. These include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. The data was stored in a CSV file and it contains some missing values for both categorical and numerical variables. For the categorical variables the missing values were taken out of the sample data using excel spreadsheet. For the numerical variables (Loan Amount), the missing values were replaced with the mean of the appropriate fields using IBM SPSS Statistics. The values for some variables were in text format and not appropriate for analysis, therefore, they were recoded into new variables for the analysis. After organizing the data, we began the analysis.

**Frequency Analysis**

The aim of this analysis was to determine the distribution of the categorical variables given in the data. This was to provide a clear picture of how the data is distributed across the different categories and is important in identifying dominant categories.

***Results and discussion***

Based on a given sample of 499 observations, Table 1 shows the frequency analysis results based on the given categorical variables: Gender, Married, Education, Self\_Employed, Property\_Area, Credit History and Loan\_Status. The frequency analysis conducted shows that majority of the loan applications were completed by Males representing 82.4% while the female applicants make up only 17.6%. Generally speaking, cultural and societal norms dictate that it is the responsibility of Men to provide shelter for their family rather than females. Adult Males strive to acquire homes even when without a family in anticipation of such. Therefore, it is understandable that Males represent the majority in Home loan applications. For the Married variable, the majority of loan applicants were married representing 64.9% while the unmarried loan applicants accounted for 35.1%. In addition to this, more loan applicants were Graduates representing 79.2% whereas 20.8% were not graduates. 86.2% of all loan applicants were not Self Employed while only 13.8 were. Furthermore, the highest number of loan applications came for properties in Semi-urban areas followed by Urban and then Rural areas with 39.1%, 31.9% and 29.1% respectively. 85.2% of home loan applicants had credit history while only 14.8% did not. Finally, 68.3% of home loan applicants had their loans approved whereas 31.7% had a “No” on their loan status.

**Table 1:** Frequency analysis results

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variables** | **Frequency** | **Percent** |
| Gender | Male | 411 | 82.4 |
|  | Female | 88 | 17.6 |
|  | *Total* | *499* | *100* |
| Married | No | 175 | 35.1 |
|  | Yes | 324 | 64.9 |
|  | *Total* | *499* | *100* |
| Education | Not Graduate | 104 | 20.8 |
|  | Graduate | 395 | 79.2 |
|  | *Total* | *499* | *100* |
| Self Employed | No | 430 | 86.2 |
|  | Yes | 69 | 13.8 |
|  | *Total* | *499* | *100* |
| Property Area | Rural | 145 | 29.1 |
|  | Urban | 159 | 31.9 |
|  | Semiurban | 195 | 39.1 |
|  | *Total* | *499* | *100* |
| Credit History | No | 74 | 14.8 |
|  | Yes | 425 | 85.2 |
|  | *Total* | *499* | *100* |
| Loan Status | No | 158 | 31.7 |
|  | Yes | 341 | 68.3 |
|  | *Total* | *499* | *100* |

**Descriptive Analysis**

The objective of this analysis was to investigate and summarize the main characteristics of the numerical variables in the data. This analysis will enable us examine the measures of central tendency (mean) and the variability of the data (standard deviation).

***Results and discussion***

Based on a sample size of 499 home loan applications, table 2 provides the results of the descriptive statistics for the variables Applicant Income, Coapplicant income, Number of Dependents, Loan amount and Loan amount term. The analysis results indicate that the mean value of an applicant’s Income was $5336.13±5618.157, mean value of a Co-applicants income was 1566.64±2580.955, mean value of the number of dependents an applicant has is 0.78±1.022, mean value of the loan amount requested was 144.735.0±78.957 and the mean value of loan amount term is 342.01 months ±64.86 months. The variability of Applicant Income was the highest around its average suggesting a wider spread of income among the loan applicants while the variability of dependents was the lowest suggesting a smaller spread of dependents among the loan applicants.

**Table 2:** Descriptive Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **N** | **Minimum** | **Maximum** | **Mean** | **Std. Deviation** |
| ApplicantIncome | 499 | 150 | 81000 | 5336.13 | 5618.16 |
| Coapplicant Income | 499 | 0 | 33837.0 | 1567.00 | 2580.95 |
| Dependents | 499 | 0 | 3 | 0.78 | 1.02 |
| Loan Amount | 499 | 9.0 | 600.0 | 144.74 | 78.96 |
| Loan Amount Term | 499 | 36 | 480 | 342.01 | 64.86 |

**Correlation Analysis**

The aim of this study was to Investigate whether there was a Linear relationship between some Numerical variables in the data set. We examine the linear relationship between each pair of ApplicantIncome, Loan\_Amount, and Loan\_term of the 499 applicants.

H0: (The variables X and Y are independent)

HA: (The variables X and Y are not independent)

***Results and discussion***

This study sought to investigate whether there is a linear correlation between each pair of the given numerical variables ApplicantIncome, Loan\_Amount and Loan\_Amount\_Term. This analysis is based on a sample size of 499 observations. First of all, the Shapiro-Wilk test was run to determine whether the variables satisfy the normality assumption. The results of the Shapiro-Wilk test asserts that the data is not normally distributed. Therefore, the Spearman Rank-order test is utilized for the analysis.

The results of the Spearman-Rank-order test are provided in Table 3. According to the results, there is a positive, moderate and a statistically significant relationship between ApplicantIncome and Loan\_Amount. (r=0.504; p-value <0.001), there is a negative, weak and not statistically significant relationship between ApplicantIncome and Loan\_Amount\_Term. (r= -0.038; p-value =0.398), there is a positive, weak and not statistically significant relationship between Loan\_Amount and Loan\_Amount\_Terms. (r=0.036; p-value =0.422).

**Table 3:** Spearman’s Correlation analysis (n=499)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **1** | **2** | **3** |
| ApplicantIncome | 1 |  |  |
| LoanAmount | 0.504\*\*\* | 1 |  |
| Loan\_Amount\_Term | -0.038 | 0.036 | 1 |
| *\*\*\* p-value<0.001* | | |  |

**Cross table with Chi-square test**

The objective of this analysis is to use a Chi-Square Test of Independence to reveal whether there is a statistically significant association between Loan Status (Approved or Rejected) and Property Area (Urban, Semi-Urban or Rural) so that we might introduce Area specific strategies such as enhanced credit education programs and boosting overall loan portfolio performance.

H0: The variables Loan\_Status and Property\_Area are independent

HA: The variables Loan\_Status and Property\_Area are not independent

***Results and Discussion***

In this analysis we investigated the relationship between Loan\_Status and Property\_Area using the Chi-square test. Table 4 outlines the results of the analysis. Per the results outlined in the table, 68.30% of the applicants had their home loans approved where as 31.7% had their home loan applications rejected. The majority of the applications were for properties that fall in the Semi-urban areas while the least number of home loan applications were received for properties in the Rural areas with 39.10% and 29.10% respectively.

According to the results, the highest rejection of 36.70% of home loans applications for properties in the Urban areas were rejected, followed by 36.10% for properties in the Rural areas and 27.20% in the Semi-urban areas. For the home loan applications that got approved, 44.60% of them are in the Semi-urban areas where as 25.80% of approvals were for properties in the rural areas. The results suggest that a loan application for a property in a Semi-urban area is more likely to be approved or than an application for a home loan for a property in the Rural area. This might be due to the fact that there is a lot of infrastructure developments ongoing in Semi-urban area coupled with the fact that Semi-urban areas are less expensive than Urban areas attracts more people especially the middle class to want to own homes in these areas.

Finally the results suggests that there is a statistically significant association between Loan\_Status and Property\_Area (χ2= 13.948, d.f : 2, p-value<0.001). Since the p-value of the analysis is less than alpha = 0.05, the null hypothesis is rejected in favor of the alternate hypothesis.

**Table 4:** Cross tabular analysis between Loan\_Status and Property\_Area

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Property\_Area** | | |  |
| **Loan\_Status** | **Rural** | **Urban** | **Semiurban** | **Total** |
| No | 57 | 58 | 43 | 158 |
|  | 36.10% | 36.70% | 27.20% | 100.00% |
|  | 39.30% | 36.50% | 22.10% | 31.70% |
| Yes | 88 | 101 | 152 | 341 |
|  | 25.80% | 29.60% | 44.60% | 100.00% |
|  | 60.70% | 63.50% | 77.90% | 68.30% |
| Total | 145 | 159 | 195 | 499 |
|  | 29.10% | 31.90% | 39.10% | 100.00% |
|  | 100.00% | 100.00% | 100.00% | 100.00% |

*Note Chi-square (χ2): 13.948, d.f.:2, p-value< 0.001*

**Comparison test**

In this analysis we tested whether the mean loan amount was the same for married and unmarried applicants. We sought to understand if the mean loan amount for married applicant was significantly higher than that of unmarried applicants. This is because traditionally married people may require more space as such may require larger loan amounts than unmarried applicants.

H0: μMarried = μUnmarried (The mean Loan\_Amount for Married applicants is not different from the mean Loan\_Amount for Unmarried applicants.)

HA: μMarried ≠ μUnmarried (The mean Loan\_Amount for Married applicants is different from the mean Loan\_Amount for Unmarried applicants.)

***Results and discussion***

According to the Kolmogorov-Smirnov/ Shapiro-Wilk test for normality, it was revealed that Loan\_Amount was not normally distributed for both Married and Unmarried home loan applicants. Secondly, as determined by Levene's test (F-stat = 4.516, p-value = 0.034), the results indicated that the variance was homogeneous. Since the normality assumption was not satisfied to run a parametric test, it was most appropriate to run a non-parametric test in Mann-Whitney U on the data. The results submit that the average rank Loan\_Amount for the unmarried applicant 210.93 was significantly different than the mean rank Loan\_Amount for the married applicant 271.10. The test results submit that there is a higher and statistically significant average Loan\_Amount for married applicants than the unmarried applicant (MWU = 21513.500; p-value < 0.001). Therefore, the null hypothesis (H0: µManufacturing = µService) was rejected in support of the alternative hypothesis (HA: µManufacturing ≠ µService) since p-value = 0.001 < 0.05.

**Multiple regression Analysis**

The aim of this analysis was to determine the effect of Dependents, Credit\_History, ApplicantIncome, Co-Applicant\_Income, Gender, Property\_Area and Loan\_Amount\_Term on the Loan Amount. Given that the higher the loan amount , the higher it’s returns Dream housing Finance will use the results of this analysis to advance a targeted marketing campaign.

***Results and discussion***

In this analysis we investigated the impact of the Dependents, Credit\_History, ApplicantIncome, Co-Applicant\_Income, Gender, Property\_Area and Loan\_Amount\_Term variables on the Loan\_Amount. Given the sample size of 499 observations, a multiple regression analysis was the most appropriate analysis tool to be deployed since the dependent variable Loan\_Amount is a continuous numerical variable. Credit\_History is a binary variable coded as 1 if the home loan applicant has previously taken any loan and 0 if they have not. Gender is also a binary variable coded 1 for female and 0 for male. Additionally, for the purpose of the multiple regression analysis two dummy variables were created from the Property\_Area variable since it three categories such as Rural, Urban and Semi-urban. Table 5 contains the findings.

Given that the VIF score varied from 1.010 to 1.449, there was no multicollinearity among the variables. The VIF range falls beneath the cut-off point 10 for multicollinearity. The F-statistic = 29.364 and p-value < 0.001 indicated that the independent variables could accurately predict the Loan\_Amount, indicating that the overall model was significant. Furthurmore, 31.3% of the variance in the dependent variable Loan\_Amount is explained by the variance in Dependents, Credit\_History, ApplicantIncome, Co-Applicant\_Income, Gender, Property\_Area and Loan\_Amount\_Term (Adj-R2 =0.313).

Based on the results, the number of dependents of a loan applicant has a positive and statistically significant impact on the Loan\_Amount(β = 8.600, p-value = 0.004). The results also indicated that for Credit\_History, an applicant who has previously taken out a loan requests a Loan\_Amount 2.712 times lesser than an applicant who has never taken a home loan. However, the impact is not significant (β = -2.712, p-value = 0.744). For the ApplicantIncome, there is a positive and significant relationship with Loan\_Amount (β = 0.007, p-value < 0.001). For the Co-ApplicantIncome, there is a positive and significant relationship with Loan\_Amount (β = 0.007, p-value < 0.001). For Gender, the loan\_Amount for a Female is 6.121 units lesser than that of Males but the impact is not statistically significant on the Loan\_Amount (β = -6.121, p-value = 0.446). The Loan\_Amount\_Terms has a positive but not significant association with Loan\_Amount (β = 0.079, p-value = 0.085). The results also indicate that, Loan\_Amounts for properties in the Urban area are 14.387 points lesser than properties in the Rural areas. This might be because urban areas already have buildings and loans may be for renovations rather than completely new builds. However, the relationship is not significant (β = -14.387, p-value = 0.057). The results also indicate that Loan\_Amounts for properties in the Semi-urban areas are 4.418 points lesser than properties in the Rural areas. The relationship is not statistically significant (β = -4.418, p-value = 0.541.)

**Table 5:** Regression Analysis

|  |  |
| --- | --- |
| **Independent Variables** | **Model**  **Dependent Variable: Loan\_Amount** |
| Dependents | 8.600\*\*  (2.901) |
| Credit History  *(Yes: 1; No: 0)* | -2.712  (-0.327) |
| Applicant Income | 0.007\*\*\*  (13.189) |
| Co-applicant Income | 0.007\*\*\*  (6.444) |
| Gender  *(Female: 1; Male: 0)* | -6.121  (-0.762) |
| Loan Amount Term | 0.079  (1.724) |
| Property Area  *(Reference = Rural)* |  |
| Property Area Urban | -14.387  (-.1906) |
| Property Area Semi-Urban | -4.418  (-0.611) |
| Constant | 71.674\*\*\*  (3.820) |
| F-Statistic | 29.364\*\*\* |
| Adj-R2 | 0.313 |
| N | 499 |

*\*p<0.05; \*\*p<0.01; \*\*\*p<0.001*

**Logistic regression Analysis**

In this analysis we set out to investigate whether a Loan is approved or not (Loan\_Status) is subject to whether an applicant is self-employed or not, ApplicantIncome and the Loan\_amount. Traditionally, Applicants with a steady stream of income are preferred when granting loans and for most self-employed persons this may not be the case. We want to develop a model that can be reliably used to determine whether a loan should be approved based on the afore mentioned variables.

***Results and discussion***

In this analysis, we sought to predict Loan approval status using a binary logistic regression. The data was gathered on customers of Dream Housing Finance, a company that provides home loans and includes ApplicantIncome, Loan\_amount, Self\_Employed and Loan\_Staus. The results of this investigation is provided in table 6.

Based on a sample size of 499 home loan applicants, the dependent variable (Loan\_Status) of an applicant being approved was coded as 1 whereas an applicant denied was coded as 0. Of the independent variables, only Self\_Employed was a dichotomous variable. It was coded 1 for Yes if the applicant is self-employed and 0 for No the applicant is not self-employed. The remaining variables ApplicantIncome and Loan\_Amount are continuous numerical variables.

For this analysis the IBM SPSS Modeler 18.4 was utilized. Since some level data preprocessing had been done with IBM SPSS statistics, we imported the Statistics file (.SAV) on to the stream to carry out the analysis. We then filtered out the variables not required for this analysis using the filter tool. The type nugget was used to measure values and define variable roles in the analysis. The Loan\_status variable was set as the target variable while the others were set as inputs.

In order to determine the effect estimates of the variables, we decided to use the odds ratio. The results submit that ApplicantIncome, Loan\_amount, and Self\_Employed does not reliably predict the Loan\_Status (-2LR stat = 620.487: Chi-Square = 2.570, p-value = 0.463). This means that the overall model is not significant. Furthermore, the Nagelkerke R2 asserts that only 0.7% of variance in Loan\_status is explained by the variation in ApplicantIncome, Loan\_amount, Self\_Employed.

Based on the results of the analysis, the submission is that there is no difference in the the odds ratio between ApplicantIncome and Loan\_status . This implies that there is no association between ApplicantIncome and Loan\_Status. (p-value = 0.840 >0.05).

Also, at each point increase in the Loan\_Amount, the odds ratio for an applicant to be approved is expected to decrease by 0.998. This implies that increasing the Loan\_Amount was significantly associated with the increased likelihood of not getting approved. As such, There is a negative but not statistically significant relationship between Loan\_Amount and Loan\_Status (p-value = 0.230 >0.05).

Finally, the odds ratio for an applicant getting approved is 0.897 less for an applicant who is self-employed compared to an applicant who is not. What this means is that self-employed applicants are 0.897 times less likely to be approved for home loans than applicant who are not self-employed. Also, there is a negative but not statistically significant association between Self\_Employed and Loan\_status (p-value = 0.695 > 0.05).

**Table 6:** Logistic regression predicting Loan\_status

|  |  |
| --- | --- |
| **Independent variables** | **Odds Ratio** |
| Applicant Income | 1.000  (0.041) |
| Self Employed | 0.897  (0.154) |
| Loan Amount | 0.998  (1.440) |
| Constant | 2.841  (27.190) |
| n | 499 |
| -2LogLikelihood | 620.487 |
| Likelihood Ratio Test *(χ2/*df) | 2.570 / 3 |
| Nagelkerke approximation of R2 | 0.007 |

*\*p<0.05; \*\*p<0.01; \*\*\*p<0.001. Effect estimates are presented as odds ratios. Wald--statistics are in parentheses*

**Conclusion**

In conclusion, this project investigated the the Dream Housing Finance loan approval information to find significant trends and insights. This study identified key factors influencing loan approval outcomes by using descriptive and frequency analyses, correlation analyses, cross-tabulations with Chi-Square tests, comparison tests, multiple regression, and logistic regression. Our findings provide a solid foundation for automating the loan qualifying process in real-time. It is very important for Dream Housing Finance company to understand the relationships between variables such as ApplicantIncome, Credit\_history, Loan\_Amount etc.. so they can optimize the home loan approval model and enhance customer experience.

**References**

*Home loan Approval*. (2023, January 20). Kaggle. https://www.kaggle.com/datasets/prepinstaprime/home-loan-approval?select=loan\_sanction\_train.csv