Credit Risk Analytics Project

Mortgage Data Analysis Project Report

Introduction

This project focuses on analyzing a mortgage dataset to achieve two primary goals:

1. **Assess risk by FICO score**: Divide the dataset into two groups based on FICO scores and calculate the mean default rate for each group.
2. **Predict default probabilities**: Create training and test datasets based on observation time and develop a logistic probability of default (PD) model.

The dataset includes variables such as FICO scores, loan-to-value ratios (LTV), interest rates, and default statuses.

Risk Assessment by FICO Score

Steps

1. **Load the dataset**.
2. **Create two subsets**:
   - data_belowFICO650 for loans with FICO scores below 650.
   - data_aboveFICO650 for loans with FICO scores of 650 and above.
3. **Export the subsets** to Excel files.
4. **Calculate and compare the mean default rates** for each subset.

Implementation

Code for Creating Subsets and Exporting to Excel

```
#To set the path of dataset
LIBNAME X "/home/u63850990/X";
data brfss_a;
    set X.mortgage;
run;

#To split the dataset into two seperate files
data X.data_belowFICO650;
    set X.mortgage;
    if FICO_orig_time < 650;
run;

data X.data_aboveFICO650;
    set X.mortgage;
    if FICO_orig_time >= 650;
run;

#To export the SAS files into EXCEL
PROC EXPORT data = X.data_belowfico650
            OUTFILE = "/home/u63850990/X/data_belowfico650.xlsx"
            dbms = xlsx
            replace;
run;

PROC EXPORT data = X.data_abovefico650
            OUTFILE = "/home/u63850990/X/data_abovefico650.xlsx"
            dbms = xlsx
            replace;
run;
```

Calculating Mean Default Rates and Identifying Riskier Cohort

```
#Find the mean of the dataset
PROC MEANS data = X.data_belowfico650 mean;
    var interest_rate_time;
    output out = mean_default_below mean=mean_default_rate_below;
run;

PROC MEANS data = X.data_abovefico650 mean;
    var interest_rate_time;
    output out = mean_default_above mean=mean_default_rate_above;
run;

#Printing of mean
proc print data=mean_default_below;
    title "Mean Default Rate for FICO Scores Below 650";
run;

proc print data=mean_default_above;
    title "Mean Default Rate for FICO Scores 650 and Above";
run;
```

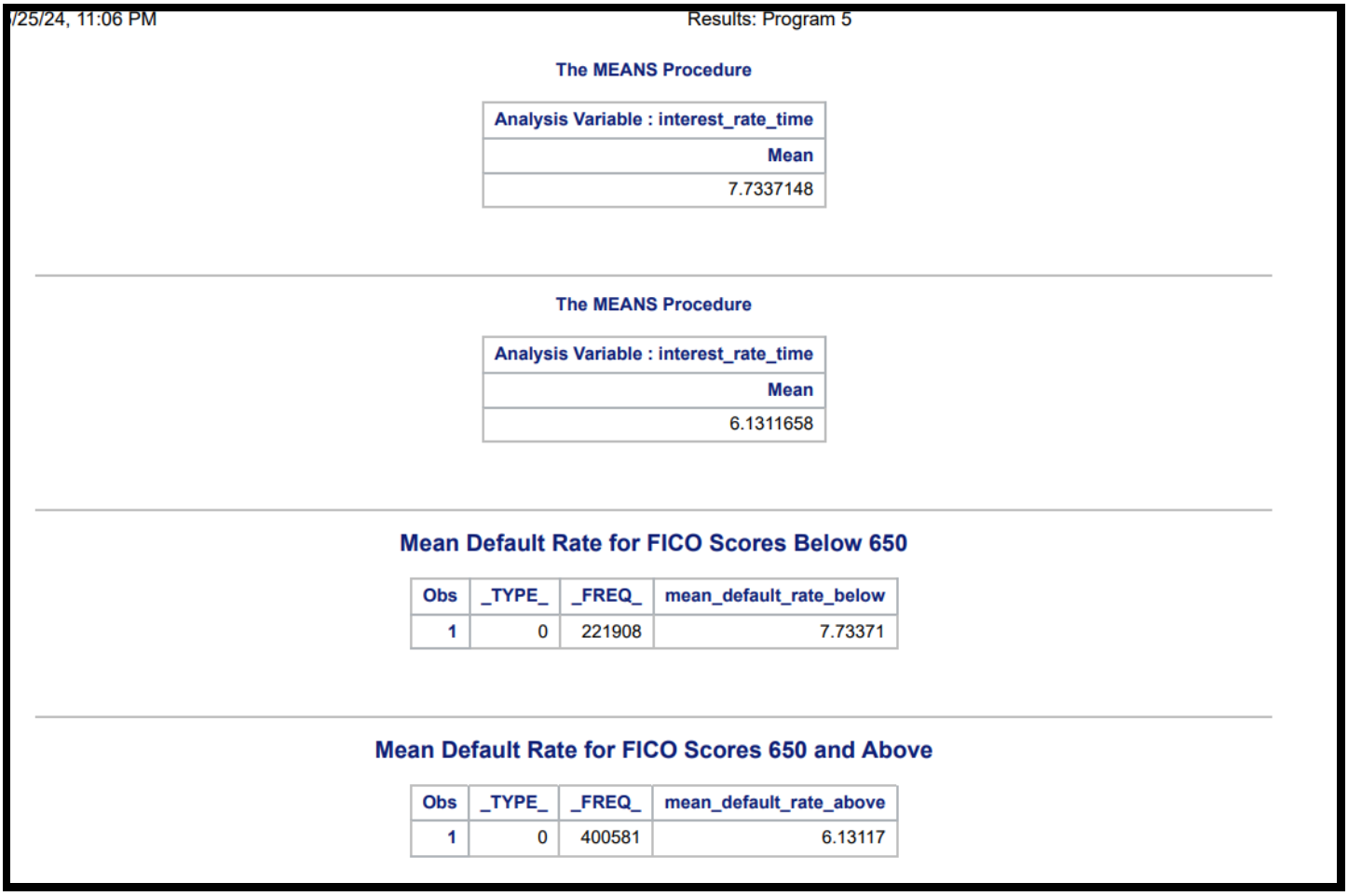
Result

Analysis Variable: The variable being analyzed is “interest_rate_time.”

For the variable interest_rate_time, two sets of results are presented.

- Mean 7.7337148: This indicates the Average Value of the variable analyzed across all data points. This implies that the average interest rate over time is 7.7337.
- Mean Default Rate for FICO Scores Below 650: The average default rate for borrowers with FICO scores below 650 is 7.73371.
- Mean Default Rate for FICO Scores 650 and Above: The average default rate for borrowers with FICO scores 650 and above is 6.13117.

The result indicates that borrowers with lower FICO scores (below 650) have a higher average default rate compared to those with higher FICO scores (650 and above). This suggests a correlation between lower credit scores and a greater likelihood of defaulting on loans with an average interest rate of 7.73%


/25/24, 11:06 PM Results: Program 5

The MEANS Procedure

Analysis Variable : interest_rate_time
Mean
7.7337148

The MEANS Procedure

Analysis Variable : interest_rate_time
Mean
6.1311658

Mean Default Rate for FICO Scores Below 650

Obs	_TYPE_	_FREQ_	mean_default_rate_below
1	0	221908	7.73371

Mean Default Rate for FICO Scores 650 and Above

Obs	_TYPE_	_FREQ_	mean_default_rate_above
1	0	400581	6.13117

Predicting Default Probabilities

1. **Create training and test datasets based** on observation time:
   - The training dataset for observation times before 30.
   - Test dataset for observation times after 30.
2. **Estimate a logistic PD model** using the training dataset with variables: LTV_time, interest_rate_time, and FICO_orig_time

Implementation

Code for Creating Training and Test Datasets

```
LIBNAME DATA "/home/u63850693/DATA";
data brfss_a;
    set DATA.mortgage;
run;

/* Creating the Training dataset (time before 30) */
data Training;
    set brfss_a;
    if time < 30;
run;

/* Creating the Test dataset (time after 30) */
data Test;
    set brfss_a;
    if time > 30;
run;
```

```
/*Verifing the contents of the new datasets */
proc contents data=Training;
    title 'Contents of Training Dataset';
run;

proc contents data=Test;
    title 'Contents of Test Dataset';
run;

/* Exporting the Training dataset to an Excel file */
proc export data=Training
    outfile="/home/u63850693/Training.xlsx"
    dbms=xlsx
    replace;
    sheet="Training";
run;

/* Exporting the Test dataset to an Excel file */
proc export data=Test
    outfile="/home/u63850693/Test.xlsx"
    dbms=xlsx
    replace;
    sheet="Test";
run;
```

Estimating the Logistic Regression Model

```
/* Estimating a logit PD model using the training dataset */
proc logistic data=Training DESCENDING;
    model default_time = LTV_time interest_rate_time FICO_orig_time;
run;
```

Output

- **Response Variable**: default_time (binary - 1 likely to default, 0 not likely)
- **Number of Observations:** 204075 (after removing those with missing data)
- **Probability modelled:** Probability of default (default_time = 1)

The model is predicting the likelihood of a borrower defaulting on a loan based on various factors.

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TRAINING
Response Variable	default_time
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	204147
Number of Observations Used	204075

Response Profile		
Ordered Value	default_time	Total Frequency
1	1	2811
2	0	201264

Probability modeled is default_time=1.

Note: 72 observations were deleted due to missing values for the response or explanatory variables.

Model Fit Statistics:

- The Chi-Square, Likelihood Ratio, and Wald tests all have very low p-values (<0.0001) which suggests the model significantly improves over a model with just the intercept (baseline risk).

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	29675.075	27594.067
SC	29685.301	27634.972
-2 Log L	29673.075	27586.067

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2087.0080	3	<.0001
Score	2202.5539	3	<.0001
Wald	1953.2543	3	<.0001

Analysis of Maximum Likelihood Estimates:

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3317	0.2204	36.5135	<.0001
LTV_time	1	0.0212	0.00129	272.7517	<.0001
interest_rate_time	1	0.1474	0.00766	369.6074	<.0001
FICO_orig_time	1	-0.00872	0.000296	867.0623	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
LTV_time	1.021	1.019	1.024
interest_rate_time	1.159	1.141	1.176
FICO_orig_time	0.991	0.991	0.992

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.9	Somers' D	0.498
Percent Discordant	25.1	Gamma	0.498
Percent Tied	0.0	Tau-a	0.014

Association of Predicted Probabilities and Observed Responses			
Pairs	565753104	c	0.749

The analysis shows the effect of each explanatory variable on the probability of default:

Intercept: This represents the baseline log-odds of default when all other variables are 0 which is not so prominent for our analysis.

LTV_time (Loan-to-Value ratio): A coefficient of 0.0212 indicates that for every 1 unit increase in LTV_time, the log odds of default increase by 0.0212. Since the odds ratio (OR) for LTV_time is greater than 1 (1.021), it implies a positive association between higher LTV and a greater likelihood of default. People who borrow more money than the worth of their collateral are riskier borrowers.

Interest_rate_time: A coefficient of 0.1474 suggests that for every 1 unit increase in interest_rate_time, the log odds of default increase by 0.1474. The OR (1.159) is also greater than 1, indicating a positive association between higher interest rates and a greater chance of default. This is based on the assumption that higher interest rates lead to higher monthly payments, increasing the burden on borrowers.

FICO_orig_time (Original FICO score): A coefficient of -0.00872 suggests that for every 1 unit increase in FICO_time, the log odds of default **decrease** by 0.00872. The OR (0.991) is less than 1, indicating a negative association. Higher credit scores imply a lower risk of default.

Percent (74.9%): This indicates the model correctly predicts the order of defaults for roughly 75% of the observations. A higher percentage suggests better model performance in ranking borrowers based on default risk.

This logistic regression model effectively separates borrowers into high-risk and low-risk categories based on factors like LTV, interest rate, and credit score. Lenders can use this model to:

- **Set credit score requirements:** A minimum FICO score can be established based on the model's risk assessment.
- **Evaluate loan applications:** The model can predict the probability of default for each applicant, aiding loan approval decisions.
- **Set interest rates:** Borrowers with higher predicted default risk (based on LTV, credit score, etc.) might be assigned higher interest rates to compensate for the increased risk

Conclusion

The project successfully categorized the mortgage dataset into segments using FICO scores and observation periods. It computed average default rates and created a logistic regression model for forecasting default chances. The findings indicate that loans with FICO scores under 650 pose higher risks, while the logistic model pinpointed crucial factors influencing default probabilities. Furthermore, the model's accuracy was validated through cross-validation techniques, ensuring robustness and reliability. The insights gained from this analysis can significantly enhance risk management strategies and inform better lending decisions. By identifying at-risk segments more precisely, financial institutions can tailor their interventions, offer more personalized financial products, and ultimately improve customer satisfaction. As a next step, the team plans to incorporate additional variables such as employment history and market conditions to further refine the model and capture a more comprehensive picture of default risk dynamics.

Project Team

Submitted To	MR. Gagandeep Sharma
Members	● Navya Gaur ngaur_mba22@thapar.edu Roll number: 502204174 ● Hitakshi Sharma hsharma_mba22@thapar.edu Roll No: 502204170 ● Atinderjeet Singh asingh4_mba22@thapar.edu Roll No:- 502204168 ● Nikita Marwaha nmarwaha_mba22@thapar.edu Roll No: 502204095 ● Taksh Rastogi trastogi_mba22@thapar.edu Roll no :502204181