**Data Analytics Capstone Topic Approval Form**

**Student Name:** Jonathan DiQuattro

**Student ID:**  000844751

**Capstone Project Name:** Multi Variable Analysis of Arkansas Mortgage Data.

**Project Topic**: Binary Regression of Arkansas Mortgage Data.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** To what extent do the independent variables predict approved loans?

**Hypothesis**: H0: There is no statistically significant relationships between the independent variables and approved loans.

H1: There is a statistically significant relationships between the independent variables and approved loans.

**Context:***.*The purpose of this study is to contribute to the field of Data Analytics by analyzing the relationship between multiple variables of the data set to gain an understanding of how and if different aspects of the data might influence the decision to approve or disapprove loan applicants. Minorities may often face an uphill battle in securing home loans across the country (DeSilver & Bialik, 2017), and this study hopes to analyze data from Arkansas state to bring awareness to current loan practices, and to give consumers confidence in applying for loans.

**Data:** The data being analyzed is made publicly available through the Consumer Financial Protection Bureau, and is distributed to fulfill the mission of the CFPB as stated on their website “The CFPB implements and enforces federal consumer financial laws to ensure that all consumers have access to markets for consumer financial products and services that are fair, transparent, and competitive” This dataset can be found at the CFPB website:

https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=ar&records=all-records&field\_descriptions=labels

The data set contains 125,392 rows of consumer mortgage applications, and consists of the following fields and types:

| **Field** | **Type** |
| --- | --- |
| as\_of\_year | Year |
| respondent\_id | Alpha-Numerical |
| agency\_name | Categorical |
| agency\_abbr | Categorical |
| agency\_code | Categorical |
| loan\_type\_name | Categorical |
| property\_type\_name | Categorical |
| property\_type | Categorical |
| loan\_purpose\_name | Categorical |
| loan\_purpose | Continuous |
| owner\_occupancy\_name | Categorical |
| loan\_amount\_000s | Numerical |
| preapproval\_name | Categorical |
| preapproval | Categorical |
| action\_taken\_name | Dependent, Categorical |
| action\_taken | Categorical |
| msamd\_name | Categorical |
| state\_name | Categorical |
| state\_abbr | Categorical |
| state\_code | Categorical |
| county\_code | Categorical |
| county\_name | Categorical |
| census\_tract\_number | Categorical |
| applicant\_ethnicity\_name | Categorical |
| applicant\_ethnicity | Categorical |
| co\_applicant\_ethnicity\_name | Categorical |
| co\_applicant\_ethnicity | Categorical |
| applicant\_race\_name\_1 | Categorical |
| applicant\_race\_1 | Categorical |
| applicant\_race\_name\_2 | Categorical |
| applicant\_race\_2 | Categorical |
| applicant\_race\_name\_3 | Categorical |
| applicant\_race\_3 | Categorical |
| applicant\_race\_name\_4 | Categorical |
| applicant\_race\_4 | Categorical |
| applicant\_race\_name\_5 | Categorical |
| applicant\_race\_5 | Categorical |
| co\_applicant\_race\_name\_1 | Categorical |
| co\_applicant\_race\_1 | Categorical |
| co\_applicant\_race\_name\_2 | Categorical |
| co\_applicant\_race\_2 | Categorical |
| co\_applicant\_race\_name\_3 | Categorical |
| co\_applicant\_race\_3 | Categorical |
| co\_applicant\_race\_name\_4 | Categorical |
| co\_applicant\_race\_4 | Categorical |
| co\_applicant\_race\_name\_5 | Categorical |
| co\_applicant\_race\_5 | Categorical |
| applicant\_sex\_name | Categorical |
| applicant\_sex | Categorical |
| co\_applicant\_sex\_name | Categorical |
| co\_applicant\_sex | Categorical |
| applicant\_income\_000s | Categorical |
| purchaser\_type\_name | Categorical |
| purchaser\_type | Categorical |
| denial\_reason\_name\_1 | Categorical |
| denial\_reason\_1 | Categorical |
| denial\_reason\_name\_2 | Categorical |
| denial\_reason\_2 | Categorical |
| denial\_reason\_name\_3 | Categorical |
| denial\_reason\_3 | Categorical |
| rate\_spread | Categorical |
| hoepa\_status\_name | Boolean |
| hoepa\_status | Categorical |
| lien\_status\_name | Categorical |
| lien\_status | Categorical |
| edit\_status\_name | Categorical |
| edit\_status | Categorical |
| sequence\_number | Numerical |
| population | Numerical, Continuous |
| minority\_population | Numerical, Continuous |
| hud\_median\_family\_income | Numerical, Continuous |
| tract\_to\_msamd\_income | Numerical, Continuous |
| number\_of\_owner\_occupied\_units | Numerical, Continuous |
| number\_of\_1\_to\_4\_family\_units | Numerical, Continuous |
| application\_date\_indicator | Date |

This data is available at the Consumer Financial Protection Bureau website and is gathered by thousands of financial institutions who report under the Home Mortgage Disclosure Act (U.S. Consumer Financial Protection Bureau, 2017). This dataset has been modified to remove any personally identifiable information, and consists of demographics, the type of loan, and the property for which the loan application is submitted for. The binary response variable (action\_taken\_name) is a categorical variable that represents if a loan was approved or denied. There are some limitations that consist of missing inputs for many of the variables, occasional errors in some inputs (for example, entering a 1 where a 10 should be). Data will need to be scrubbed to remove erroneous information and missing information. The delimitations for this study are to have enough fields with imputed data so that reliable conclusions can be drawn from the statistical analysis, and to have enough examples of denied applications and approved applications to have a wealth of target events to avoid over sampling.

**Data Gathering:** The data will be collected from the Consumer Financial Protection Bureau website as a downloadable comma separated file. This file will be analyzed with Microsoft Excel and SAS studio. The data will be cleaned using a combination of the previous mentioned technologies. Some fields will need to be converted, for example the applicant\_income will be converted from a multiple of 1000 to the actual integer, and the hoepa\_status\_name field will be converted from “yes” and “no” to a “0” or “1”. The response variable is categorical and needs to be converted from the various denied and approved explanations to “0” and “1”.

**Data Analytics Tools and Techniques**: To determine the relationship between variables in the data set and how they influence the result of getting an approved loan, a binary regression model will be created to explain how different variables in the data set influence each other and the overall result. Pearson correlation will be applied to quantify the relationships between each variable and the binary response. The dataset has a many variables and will need to be reduced via variable screening to eliminate high correlations between variables other than the target variables. This will be achieved using PROC VARCLUS which will eliminate redundancy and reduce variables.

**Justification of Tools/Techniques:** SAS is the primary tool that will be used for the binary regression and is more than capable in handling many of the issues the commonly arise when running binary regression. For the redundant and secondary relationships between variables other than the target PROC VARCLUS will be applied prior to running the logistic regression. This procedure is made very easy with SAS and will not require researching other packages that might be required with other solutions such as R or Python. The version of SAS being used for this research (SAS Studio) does have a limit to the amount of data that can be uploaded (100MB), but due to the data for this study only consisting of 97MB SAS Studio will be more than capable.

**Project Outcomes**:This project will attempt to give consumers confidence in applying for mortgages by explaining which variables are the best indicators of getting an approved loan, and provide insights on what a planning consumer should focus on to improve their chances of getting an approved loan.

**Projected Project End Date**: 10/21/2021

**Sources**:

U.S. Consumer Financial Protection Bureau. (2017). Retrieved September 19, 2021, from <https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=ar&records=all-records&field_descriptions=labels>

DeSilver, D., &; Bialik, K. (2017, January 10). Blacks, hispanics face mortgage challenges. Pew Research Center. Retrieved September 19, 2021, from [https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/](https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/.)

SAS. (n.d.). The VARCLUS Procedure. SAS help center. Retrieved September 19, 2021, from <https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_varclus_syntax01.htm>

To be filled out by a course mentor:

☐ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Mentor’s Approval Status: Select one

Date: Click here to enter a date.

Reviewed by: Click here to enter text.

Comments: Click here to enter text.