**Data Analytics Capstone Topic Approval Form**

**Student Name:**  Sample Proposal

**Student ID:**  99999999

**Capstone Project Name:** Linear Regression on PPP Loan Data

**Project Topic**: Predictive Model for PPP Loan Data

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

**Research Question:** Can a regression model be constructed based solely on the research data?

**Hypothesis**: H0: A predictive regression model cannot be made from the PPP dataset.

H1: A predictive model can be constructed from the research dataset.

**Context:***.*The contribution of this study to the field of Data Analytics and the MSDA program is to create a predictive model which can estimate company loan amounts so a company does not need to go through an extensive process of gathering all the documentation needed to apply (U.S. Department of the Treasury, 2020). This study will utilize a Generalized Linear Model to analyze the significance of predictor variables and identify which variables best predict the dollar amounts of PPP Loans. A Generalized Linear Model allows the expression of relationships between predictor variables and the response variable (Date, 2020). Reynolds (2018) found that utilizing multiple linear regression a predictive can be created. The researchers hypothesized that loan data can be descriptive enough to determine predictability of outcome. By understanding the relationship between the predictor variables and the response variable, it should be possible to predict the response variable given the predictor inputs (Statistical Consulting Group, 2020).

**Data:** The data needed to be collected for the question is the publicly available information provided by the Small Business Administration for Paycheck Protection Program loan recipients in the state of New York (U.S. Department of the Treasury, 2020). The data set before removing these records contains 301,796 rows.

The data set is made available through the Small Business Administration. The data set includes the following variables of zip code, naics code, business type, raceethnicity, gender, veteran, non-profit, jobsreported, dateapproved, lender and the target variable loan amount. The predictor variables are broken down as follows:

<https://home.treasury.gov/policy-issues/cares-act/assistance-for-small-businesses/sba-paycheck-protection-program-loan-level-data>

|  |  |
| --- | --- |
| **Field** | **Type** |
| City | Categorical |
| State | Categorical |
| Zip | Categorical |
| NAICSCode | Categorical |
| BusinessType | Categorical |
| RaceEthnicity | Categorical |
| Gender | Categorical |
| Veteran | Categorical |
| NonProfit | Categorical |
| JobsReported | Continuous |
| DateApproved | Categorical |
| Lender | Categorical |
| Congressional District | Categorical |

The Small Business Administration has made this information publicly available. There is no information that would make the companies approved for these loans identifiable. Limitations: The data set is limited by the accuracy and completeness of information made publicly available by the SBA and only contains loans below $150,000. Additionally, as all fields were not required to be answered by loan applicants, there is a significant amount of missing inputs regarding Gender, RaceEthnicity and Veteran status. Delimitations: The data set will be delimited by removing records which have missing information in the JobsReported field. Additionally the variable of State will be removed as all records are in the State of New York. It is important to have a sufficient number of rows to build a properly fitted Linear Regression Model (Austin & Steyerberg, 2015).

**Data Gathering:** The Treatment of the Data: Data will be downloaded a publicly available CSV file from the sba.gov website which shows data for all PPP loans under $150,000 issued in the State of New York. Any entries that have missing inputs for the variable of JobsReported will be removed. Failure to remove the missing inputs can negatively impact the accuracy of the model (Swalin, 2018). The data quality is very high as the data has been collected and distributed by the Small Business Administration. The data contains both Quantitative and Qualitative variables. SAS will be utilized to clean the data and remove missing variables or anomalies. All categorical variables will be converted to binary variables with the creation of dummy variables. Further missing values will be imputed. Overall data sparsity is < 10%.

**Data Analytics Tools and Techniques**: The Design of the Study: 1. A Q-Q Plot and Shapiro-Wilk were run to determine normality of the data. 2. A Generalized Linear Model which does not require a test for normality (University of Colorado Boulder, 2018) nevertheless will be run for further corroboration. The Generalized Linear Model will utilize stepwise analysis to identify the contribution of each independent variable in predicting the dependent variable. During the process of model fitting, a portion of the data will be fragmented for honest assessment of the model. The process of fitting a Generalized Linear Model should create a model which can accurately predict the value of the dependent variable given the independent variable inputs (Date, 2020). Subsequent cluster analysis will be performed to identify additional insights into the data segmentation.

**Justification of Tools/Techniques:** SAS will be used for the creation of the regression model for the data. According to Kunal (2017) SAS is stronger than R or Python in the ability to clean and analyze large data sets. Additionally SAS is preferred over Python and R for regression analysis (Kunal, 2017).

**Project Outcomes**: The project will seek to create a linear regression model for the dollar amount of PPP loans based on the composition of the applying company. Support for the alternative hypothesis is found in Martinson (2020) that a linear regression model can be helpful in estimating the target variable by determining the strength of predictor variables with respect to loan data.

**Projected Project End Date**: 10/24/2020

**Sources**:

U.S. Department of the Treasury. (2020, October 11). Retrieved October 13, 2020, from <https://home.treasury.gov/policy-issues/cares-act/assistance-for-small-businesses/sba-paycheck-protection-program-loan-level-data>

How to Complete Your PPP Loan Application. (n.d.). Retrieved October 13, 2020, from <https://www.sba.com/funding-a-business/government-small-business-loans/ppp/how-to-complete-paycheck-protection-program/>

Austin, P., & ; Steyerberg, E. (2015, January 22). The number of subjects per variable required in linear regression analyses. Retrieved October 13, 2020, from <https://www.sciencedirect.com/science/article/pii/S0895435615000141>

Generalized linear model. (2020, September 26). Retrieved October 13, 2020, from <https://en.wikipedia.org/wiki/Generalized_linear_model>

JainKunal, K. (2020, June 07). Python vs R vs SAS: Which Data Analysis Tool should I Learn? Retrieved October 13, 2020, from <https://www.analyticsvidhya.com/blog/2017/09/sas-vs-vs-python-tool-learn/>

What is Linear Regression? (n.d.). Retrieved October 13, 2020, from <https://www.statisticssolutions.com/what-is-linear-regression/>

Date, S. (2020, May 10). Generalized Linear Models. Retrieved October 13, 2020, from <https://towardsdatascience.com/generalized-linear-models-9ec4dfe3dc3f>

Coding Systems for Categorical Variables in Regression Analysis. (n.d.). Retrieved October 22, 2020, from https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis-2/

Analyzing Non-Normal Data with Generalized Linear Models (GLMs). (2018). Retrieved October 18, 2020, from <https://www.colorado.edu/lab/lisa/services/short-courses/analyzing-non-normal-data-generalized-linear-models-glms>

Swalin, A. (2018, March 19). How to Handle Missing Data. Retrieved October 13, 2020, from <https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>

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.