**Matt Purvis – BUA 6110 – Predictive & Prescriptive Analytics – Bank Customer Logistic Regression DDD**

Table of Contents

[DDD Summary Steps 2](#_Toc69847384)

[Character Variables 2](#_Toc69847385)

[Missing Values 2](#_Toc69847386)

[Outliers 3](#_Toc69847387)

[Relationships 4](#_Toc69847388)

[Model Build 7](#_Toc69847389)

[Dataset 7](#_Toc69847390)

[Instructions and Deliverables for Logistic Regression Model 8](#_Toc69847391)

[Initial Model Run 8](#_Toc69847392)

[Correlations 8](#_Toc69847393)

[Diagnostic Plot 9](#_Toc69847394)

[Significant and Non-Significant Variables 9](#_Toc69847395)

[Variables Review 10](#_Toc69847396)

[Second Model Run 11](#_Toc69847397)

[Correlations 11](#_Toc69847398)

[Diagnostic Plot 11](#_Toc69847399)

[Significant and Non-Significant Variables 12](#_Toc69847400)

[Variables Review 12](#_Toc69847401)

[Third Model Run 13](#_Toc69847402)

[Correlations 13](#_Toc69847403)

[Diagnostic Plot 13](#_Toc69847404)

[Significant and Non-Significant Variables 14](#_Toc69847405)

[Variables Review 14](#_Toc69847406)

[Final Model Results – Evaluating the Best Model 15](#_Toc69847407)

# DDD Summary Steps

## **Character Variables**

1. Are there any character variables? If yes, recode character variables to numeric variables so you can use them for subsequent data exploration and model build

Yes, there were several binary character variables that had either a 0 or 1 input for the value, which signified Yes (1) or No (0). For the Gender\_P1 column, Male = 0 and Female = 1. Below is a list of the character variables used in this analysis.

|  |  |
| --- | --- |
| Binary Variables | |
| Checking\_flag | Auto\_Refinance |
| Savings\_flag | Credit\_Card |
| Loans\_LOC\_flag | Home\_Equity\_ITA |
| CreditCard\_flag | Click |
| CDs\_flag | Open |
| HELoans\_HELOC\_flag | Has\_kids |
| MMDA\_flag | Multi\_Adult |
| Gender\_P1 |  |

## **Missing Values**

1. Are there any missing values? What is the % of customers with missing value for each predictor. For predictor with missing values, describe and show how you would treat them.

|  |  |  |
| --- | --- | --- |
| Variable | # Missing | Action |
| MOBILE\_ACTIVE | 2997 | Binary; converted nulls to 0 |
| OLB\_ACTIVE | 1311 | Binary; converted nulls to 0 |
| OLB\_ENROLLED | 640 | Binary; converted nulls to 0 |
| Checking\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| Savings\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| Loans\_LOC\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| CreditCard\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| CDs\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| HELoans\_HELOC\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| MMDA\_avg\_Bal | 21 | Filtered out of dataset due to small number of missing values |
| signon\_trans\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| signon\_visit\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| Branch\_Visits\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| Branch\_Trans\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| ATMVisits\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| ATM\_Trans\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| annual\_revenue | 8 | Filtered out of dataset due to small number of missing values |
| CC\_Visits\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| VRU\_visits\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| BP\_trans\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |
| RDC\_trans\_3\_Mos | 8 | Filtered out of dataset due to small number of missing values |

It is also worth noting that there were other missing variable actions taken for the 3rd party dataset . Please see the ‘.HTML’ file attached with my submission to see explanation of treatment of additional missing values.

## **Outliers**

1. Are there any outliers? Show how you detect outliers and how you’d treat outliers.

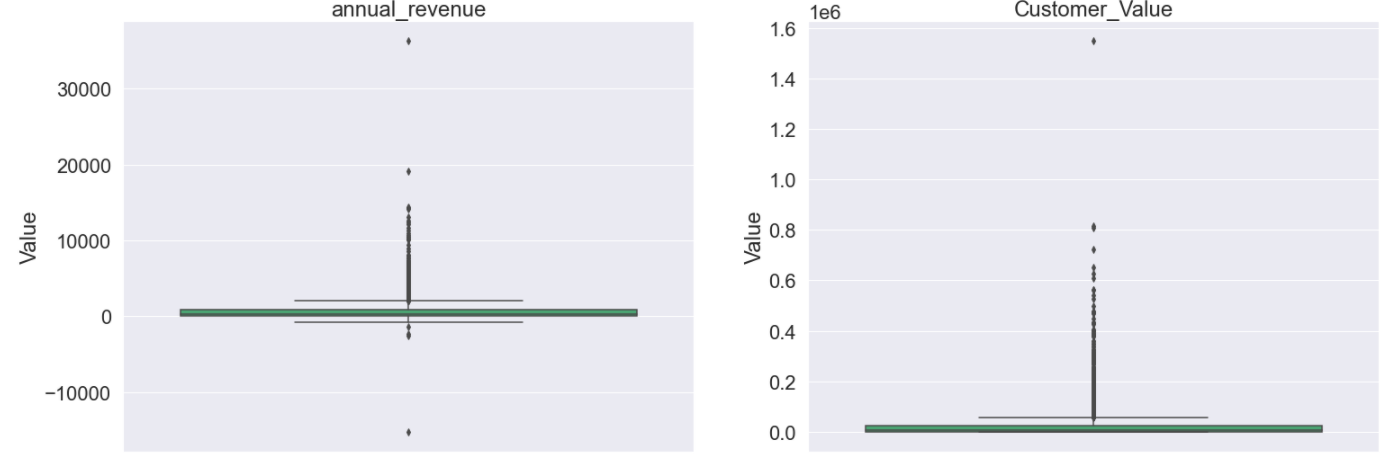
Outliers were assessed for both the ‘annual revenue’ and ‘Customer Value’ fields. First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

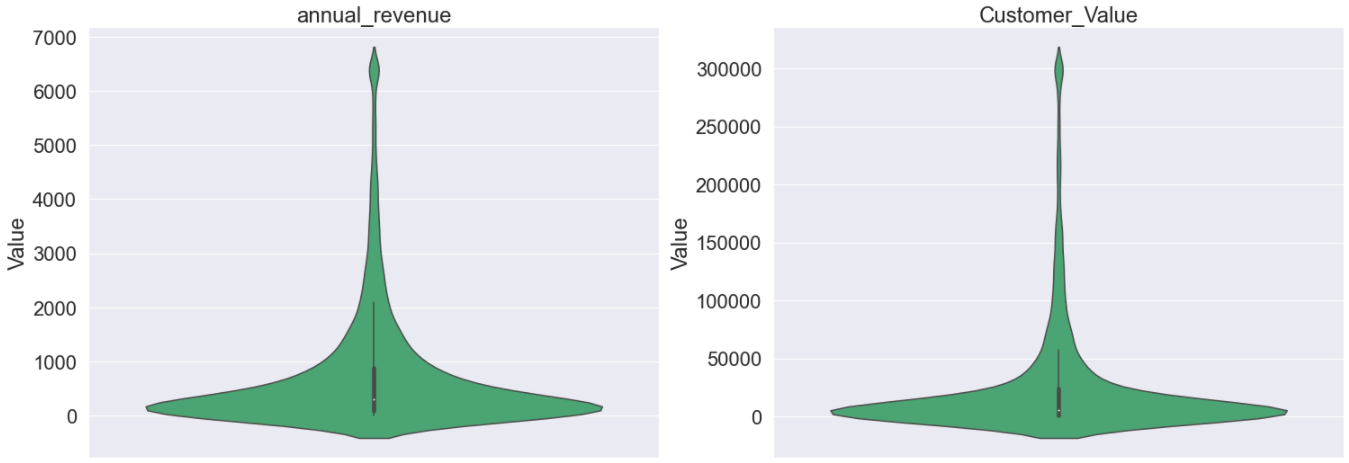
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in annual\_revenue is: 6374
  + The 1st percentile in annual\_revenue is: -40
  + The 99th percentile in Customer\_Value is: 296,261.
  + The 1st percentile in Customer\_Value is: 0
* Annual Revenue:
  + Ceiling: Use 99th percentile at 6400
  + Floor: Use 0 as the lowest allowable value
* Customer\_Value columns:
  + Ceiling: Use a number approximating the 99th percentile ($300,00)
  + Floor: Use 0 as the lowest allowable value

The following page shows the before and after. Please note that I use a box plot for the *before visual* and a violin plot for the *after visual*. The reason for this is to assess the final distributions of the numeric variables.

**Before:**



**After:**



## **Relationships**

1. What are the relationship and strength of the relationship between each predictor and the dependent variable? Show supporting data to answer this question

Due to the sheer size of the visuals, please see the ‘*visuals’* folder included in the supplemental materials provided with this submission. The visuals are in three files:

* Multi Value visuals – These are for categorical variables with more than 2 values
* Binary Value visuals – These are visuals for binary variables
* Continuous variable visuals – These are visuals (box plots) for the continuous variables

## **Variable Transformations**

1. For subsequent LOGISTIC regression model build with the binary DV, show how you would recode / transform your predictors to support a robust logistic regression model build.

I will be taking the percent of total of the dependent variable (‘DV’) and grouping selected variables for a more condense and effective way to reduce the number of dummy variables. This should also help group like variables together for simplicity when interpreting the model. The variables include:

* Occupation
* Dwelling\_Type
* Education

In the following ‘Instructions and Deliverables’ section, see the Second Model run portion for further details. I will be running the initial model run with no transformations and the secondary model run will incorporate these transformations.

## **Dataset**

1. Once you completed all the steps above, save your data with all the treated variables you created. You will use this new dataset for the logistic regression model build next week

See the ‘.HTML’ code (last line) for final data output to begin more robust regression model building that will continue into the next section of this report.

# Instructions and Deliverables for Logistic Regression Model (Week 6 – Not Completed)

Use data in the “Customer” tab and define customer digital channel usage as the binary dependent variable of the logistic regression model you will build

* DV =1, if OLB\_ACTIVE=1 or MOBILE\_ACTIVE=1; else DV=0
* This DV indicates whether customers are currently active in using online banking or mobile banking

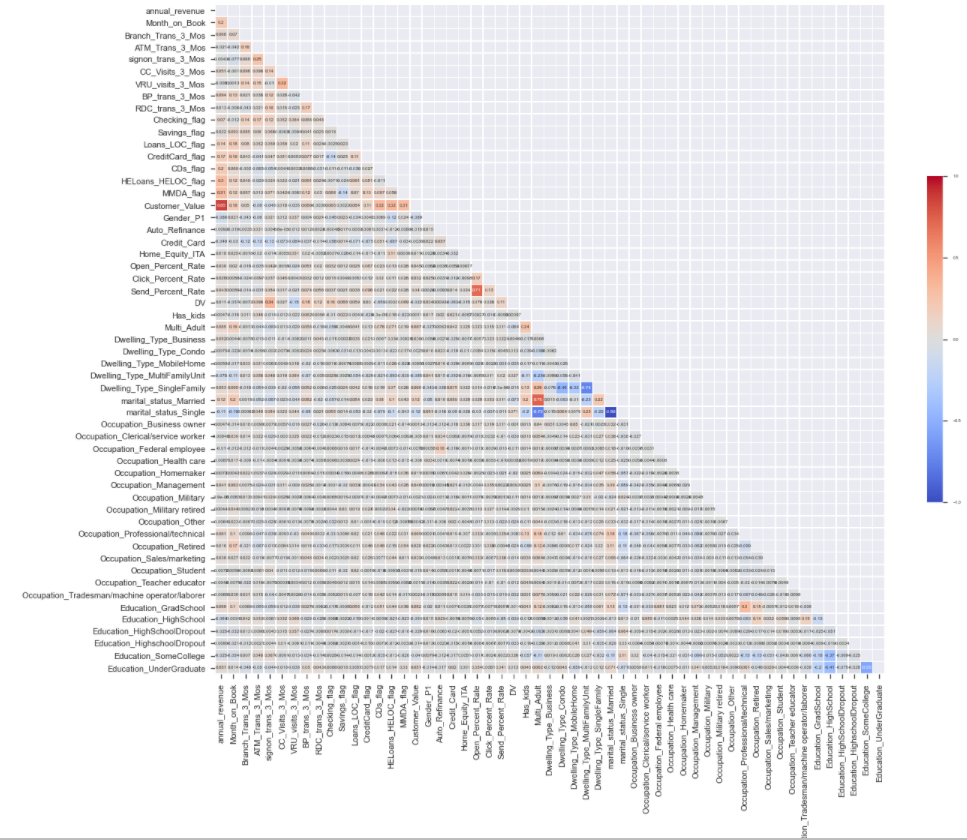
Use logistic regression technique for your model build

Run a model with all your treated predictors. Review your model results and address the following questions -

# Initial Model Run

## **Correlations**

Show correlation of each predictor with DV (i.e., univariate relationship)



## **Diagnostic Plot**

Show diagnostic plots of your predictors

