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

Table of Contents

[DDD Summary Steps 2](#_Toc69671536)

[Character Variables 2](#_Toc69671537)

[Missing Values 2](#_Toc69671538)

[Outliers 3](#_Toc69671539)

[Relationships 4](#_Toc69671540)

[Model Build 7](#_Toc69671541)

[Dataset 7](#_Toc69671542)

[Instructions and Deliverables for Logistic Regression Model (Week 6 – Not Complete) 8](#_Toc69671543)

# 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.

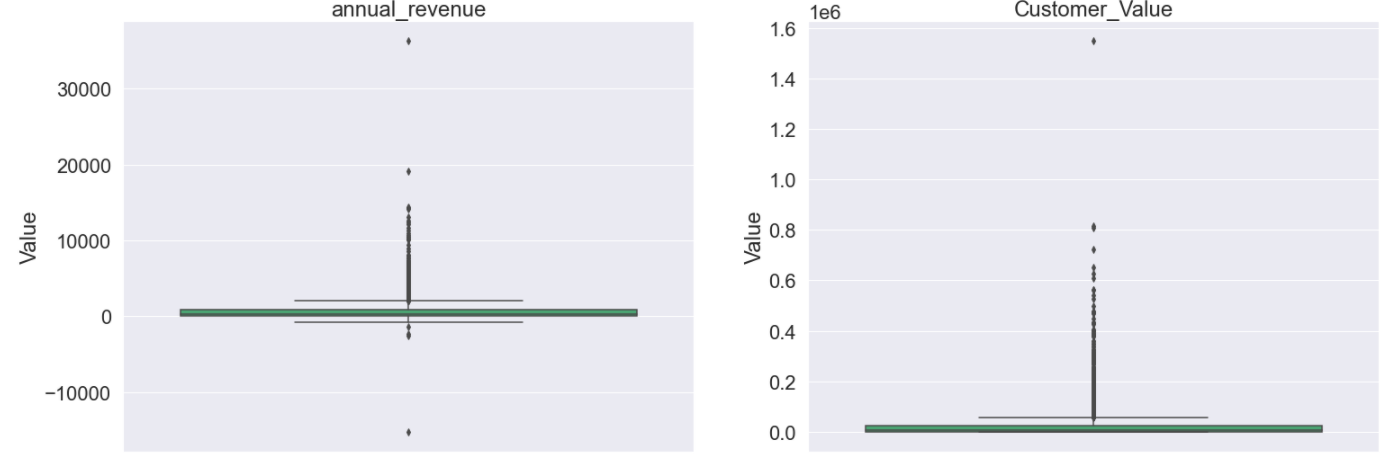
Outliers were assessed for both the ‘annual revenue’ and ‘Customer Value’ fields. First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

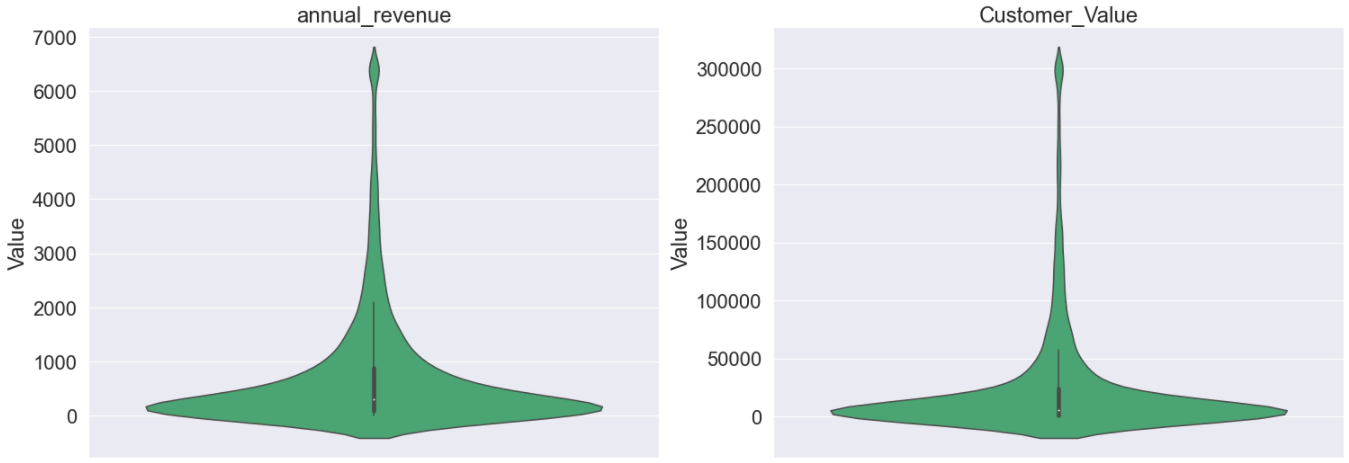
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in annual\_revenue is: 6374
  + The 1st percentile in annual\_revenue is: -40
  + The 99th percentile in Customer\_Value is: 296,261.
  + The 1st percentile in Customer\_Value is: 0
* Annual Revenue:
  + Ceiling: Use 99th 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:**

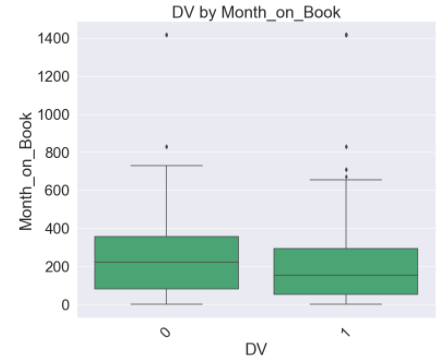


## **Relationships**

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

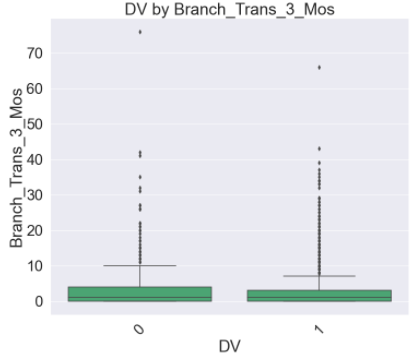
**Month\_on\_Book**

* Based on the visual below, it looks as if customers who use online or mobile banking tend to have, on average, a lower number of months with the company. However, the variability is wide for both, so this variable looks to be a limited value-add at first glance.



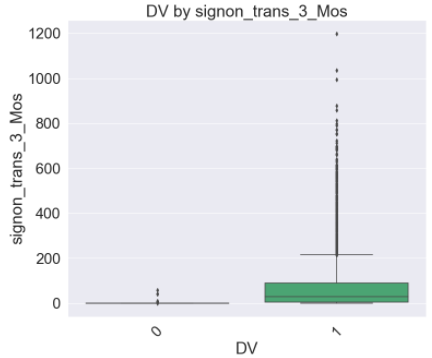
**Branch\_Trans\_3\_Mos**

* There does not look to be much of a difference between these two variables. However, it is worth continuing to better understand the variable since it has such a low p-value and can be statistically significant to predicting if a customer uses online or mobile banking.



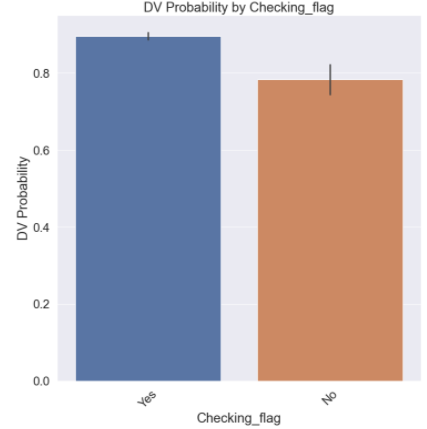
**Signon\_trans\_3\_Mos**

* This variable looks to be very important, as the number of signons should implicate if a customer has online or mobile banking. Due to its seemingly direct correlation with the dependent variable, it is worth keeping an eye on this in the event it helps with the analysis.



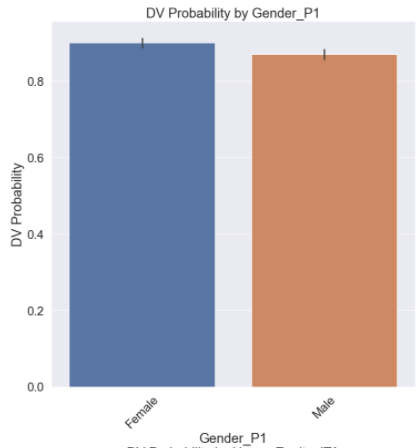
**Checking\_flag**

* The probability of a customer using online or mobile banking seems to be a bit different if a customer has a checking account.



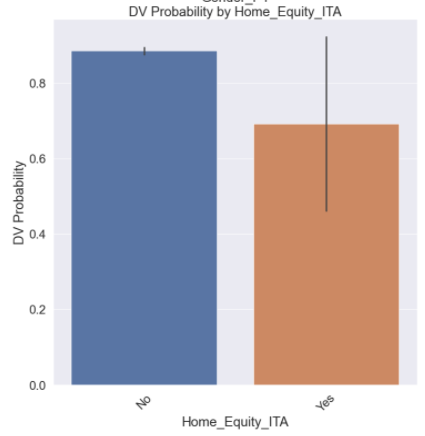
**Gender\_P1**

* There does not seem to be much variability depending on if the customer is a male or femaile. However, due to such a low p-value, it is worth exploring how important this variable is moving forward.



**Home\_Equity\_ITA**

* The probability of a customer using online or mobile banking seems to be a bit different if a customer was sent a direct mail notice regarding a Home Equity product. However, due to the low volume of these notices going to customers, I am hesitant to believe this variable will be significant as the analysis progresses.



## **Model Build**

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.

**Step 1:** Incorporate some of the 3rd party data using features such as:

* Dwelling\_Type – create dummy variables, as the variance looks to be wide-ranged for each value
* Household\_Age – There looks to be a direct correlation between the age of the head of household to the probability of a customer using online banking

**Step 2:** Transform the Numerical Variables for Scaling upon splitting the data into training and test datasets

## **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 Complete)

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 -

* Show correlation of each predictor with DV (i.e., univariate relationship)
* Show diagnostic plots of your predictors
* For your initial multiple linear regression model -3
  + What are significant variables in the model in predicting customer value? What
  + are variables that are not significant?
  + Are any variables with conflicting relationship with DV in the univariate
  + correlation analysis vs. in your multiple logistic regression model?
  + What is the impact of significant variables? Positive or negative?
  + What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

3. Remove variables that are not significant. Rerun your model and answer the same set of questions above.

4. Continue to tweak your model if needed, for example, including interaction terms, additional variable transformation, and finalize your model.

5. Comment on your final model results including

* Interpretation of the model
* Improvement of model performance from your initial model
* How will you use the model to support the Bank’s digital channel migration campaign?
* Any other comments you want to add