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

Table of Contents

[DDD Summary Steps 2](#_Toc70172433)

[Character Variables 2](#_Toc70172434)

[Missing Values 2](#_Toc70172435)

[Outliers 3](#_Toc70172436)

[Relationships 4](#_Toc70172437)

[Variable Transformations 4](#_Toc70172438)

[Dataset 4](#_Toc70172439)

[Instructions and Deliverables for Logistic Regression Model 5](#_Toc70172440)

[Initial Model Run 5](#_Toc70172441)

[Correlations 5](#_Toc70172442)

[Diagnostic Plot 6](#_Toc70172443)

[Significant and Non-Significant Variables 7](#_Toc70172444)

[Variables Review 7](#_Toc70172445)

[Second Model Run 8](#_Toc70172446)

[Correlations 8](#_Toc70172447)

[Diagnostic Plot 9](#_Toc70172448)

[Significant and Non-Significant Variables 10](#_Toc70172449)

[Variables Review 10](#_Toc70172450)

[Reiterating Diagnostic Plots 11](#_Toc70172451)

[Third Model Run 12](#_Toc70172452)

[Diagnostic Plot 12](#_Toc70172453)

[Significant and Non-Significant Variables 12](#_Toc70172454)

[Variables Review 12](#_Toc70172455)

[Final Model Results – Interpretation 13](#_Toc70172456)

[Improvement of model performance from your initial model 13](#_Toc70172457)

[Interpretation of the model 13](#_Toc70172458)

[Management Recommendations 14](#_Toc70172459)

[How will you use the model to support the Bank’s digital channel migration campaign? 14](#_Toc70172460)

# 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.

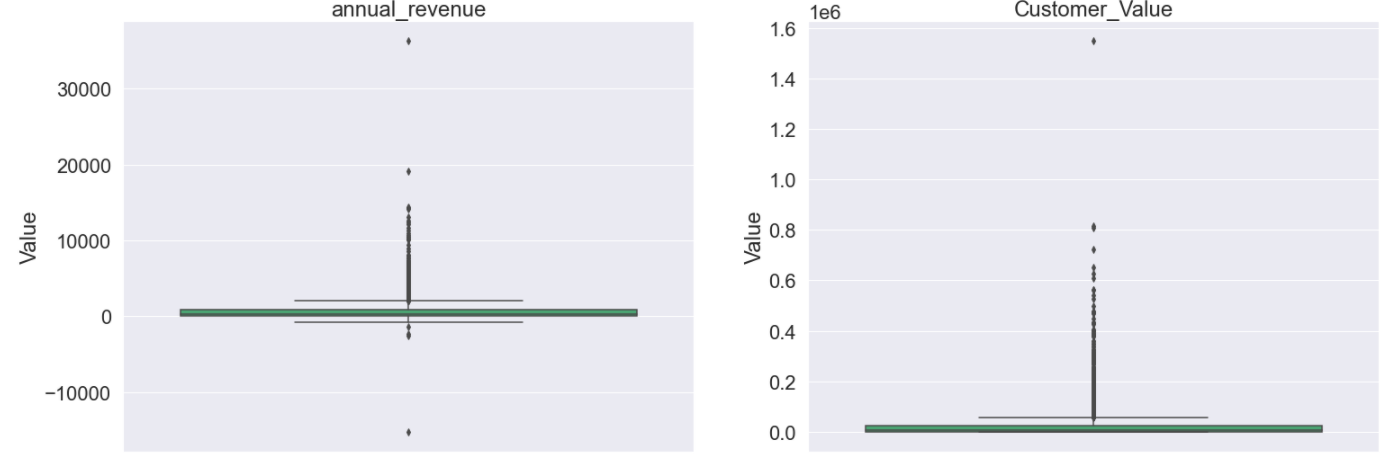
Outliers were assessed for both the ‘annual revenue’ and ‘Customer Value’ fields. First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

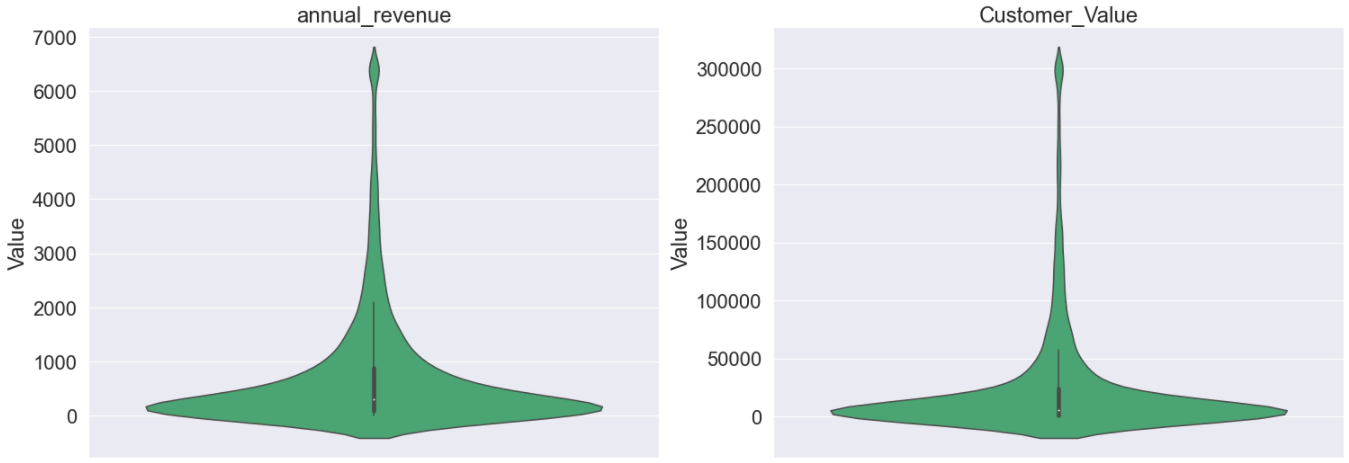
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in annual\_revenue is: 6374
  + The 1st percentile in annual\_revenue is: -40
  + The 99th percentile in Customer\_Value is: 296,261.
  + The 1st percentile in Customer\_Value is: 0
* Annual Revenue:
  + Ceiling: Use 99th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

**IMPORTANT:** After assessing the variables in the univariate analysis, I am going to remove the ‘signon\_Trans\_3\_Mos’ variable, as this variable is directly tied to the dependent DV variable. This will skew my prediction and potentially impact my overall results.

# Instructions and Deliverables for Logistic Regression Model

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

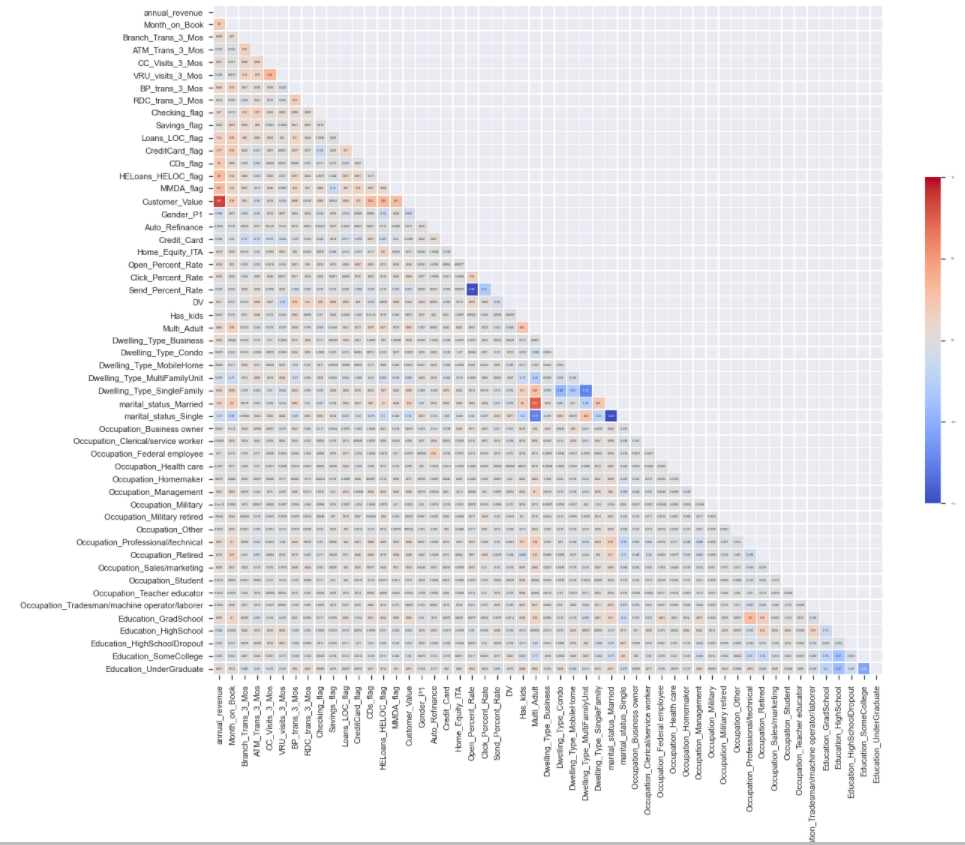
Use logistic regression technique for your model build

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

# Initial Model Run

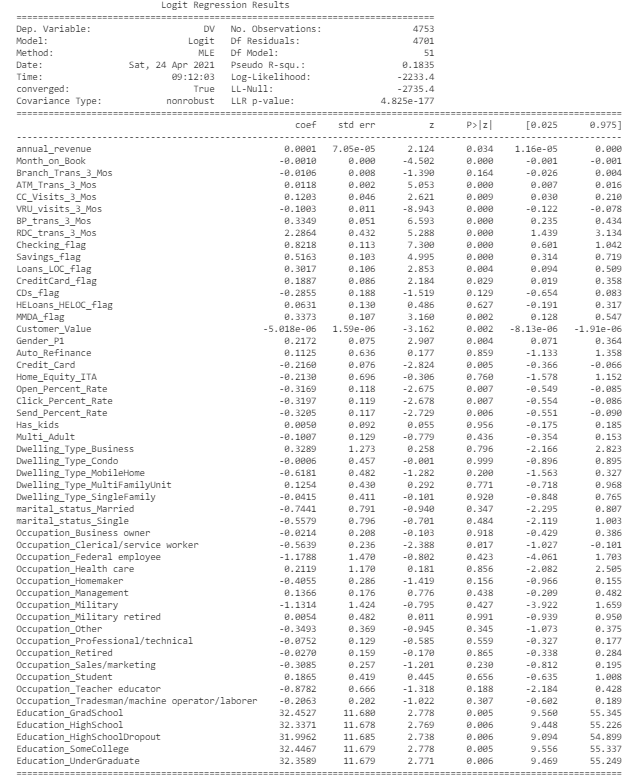
## **Correlations**

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



## **Diagnostic Plot**

Show diagnostic plots of your predictors



## **Significant and Non-Significant Variables**

What are significant variables in the model in predicting customer value? Not significant?

|  |  |
| --- | --- |
| Significant | Non-Significant |
| Annual\_revenue | Branch\_Trans\_3\_Mos |
| Month\_on\_Book | CDs\_flag |
| BP\_trans\_3\_Mos | HELoans\_HELOC\_flag |
| ATM\_Trans\_3\_Mos | Auto\_Refinance |
| CC\_Visits\_3\_Mos | Home\_Equity\_ITA |
| VRU\_visists\_3\_Mos | Has\_kids |
| RDC\_trans\_3\_Mos | Multi-Adult |
| Checking\_flag | Dwelling\_Type\_Business |
| Savings\_flag | Dwelling\_Type\_Condo |
| Loans\_LOC\_flag | Dwelling\_Type\_MobileHome |
| CreditCard\_flag | Dwelling\_Type\_MultiFamilyUnit |
| MMDA\_flag | Dwelling\_Type\_SingleFamily |
| Customer\_Value | Marital\_status\_Married |
| Gender\_P1 | Marital\_status\_Single |
| Credit\_Card | Occupation\_Business Owner |
| Open\_Percent\_Rate | Occupation\_Federal employee |
| Click\_Percent\_Rate | Occupation\_Health care |
| Send\_Percent\_Rate | Occupation\_Homemaker |
| Occupation\_Clerical/service worker | Occupation\_Management |
| Education\_GradSchool | Occupation\_Military |
| Education\_HighSchool | Occupation\_Military retired |
| Education\_HighschoolDropout | Occupation\_Other |
| Education\_HighschoolDropout | Occupation\_Professional/technical |
| Education\_SomeCollege | Occupation\_Retired |
| Education\_UnderGraduate | Occupation\_Sales/marketing |
|  | Occupation\_Student |
|  | Occupation\_Teacher educator |
|  | Occupation\_Tradesman/machineop/laborer |

*\*Red = significant p-value; negative coef*

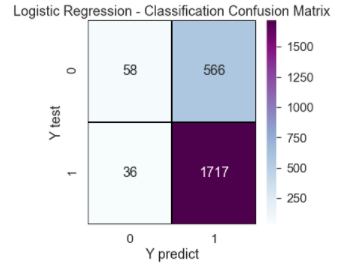
## **Variables Review**

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?

The ‘signon\_trans\_3\_Mos’ variable had a direct conflict, as it weighted too heavily against the other variables. The logic being if someone signs on to a platform, the individual/ household has access to mobile or online banking.

*See above variables marked in red to assess negatively correlated values to DV; if in black, the variables have a positive relationship.*

What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

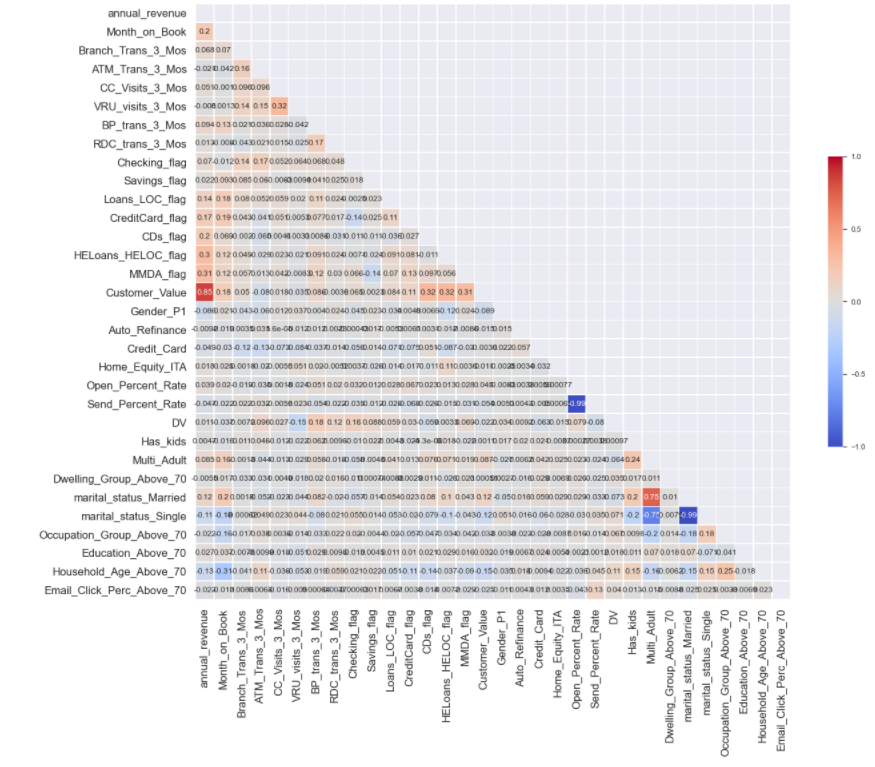
**Pseudo R-squared =** 0.1835

**Confusion Matrix:** 74.67% Accuracy 🡪

# Second Model Run

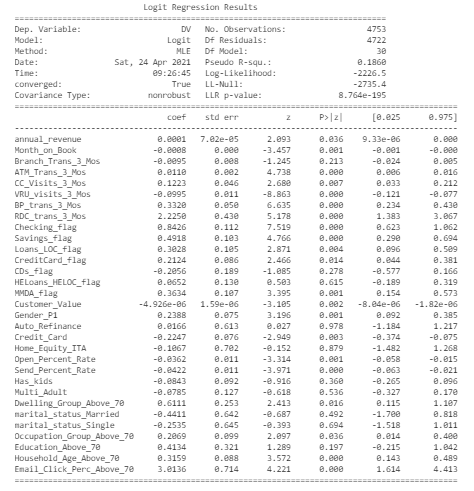
## **Correlations**

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



## **Diagnostic Plot**

Show diagnostic plots of your predictors



## **Significant and Non-Significant Variables**

What are significant variables in the model in predicting customer value? Not significant?

|  |  |
| --- | --- |
| Significant | Non-Significant |
| Annual\_revenue | Branch\_Trans\_3\_Mos |
| Month\_on\_Book | CDs\_flag |
| BP\_trans\_3\_Mos | HELoans\_HELOC\_flag |
| ATM\_Trans\_3\_Mos | Auto\_Refinance |
| CC\_Visits\_3\_Mos | Home\_Equity\_ITA |
| VRU\_visists\_3\_Mos | Has\_kids |
| RDC\_trans\_3\_Mos | Multi-Adult |
| Checking\_flag | Education\_Above\_70 |
| Savings\_flag |  |
| Loans\_LOC\_flag |  |
| CreditCard\_flag |  |
| MMDA\_flag |  |
| Customer\_Value |  |
| Gender\_P1 |  |
| Credit\_Card |  |
| Open\_Percent\_Rate |  |
| Click\_Percent\_Rate |  |
| Send\_Percent\_Rate |  |
| Dwelling\_Group\_Above\_70 |  |
| Occupation\_Group\_Above\_70 |  |
| Household\_Age\_Above\_70 |  |
| Email\_Click\_Perc\_Above\_70 |  |

*\*Red = significant p-value; negative coef*

## **Variables Review**

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?

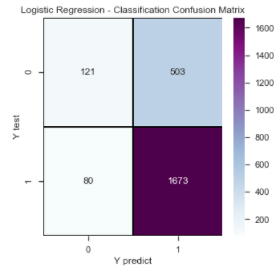
The ‘signon\_trans\_3\_Mos’ variable had a direct conflict, as it weighted too heavily against the other variables. The logic being if someone signs on to a platform, the individual/ household has access to mobile or online banking.

*See above variables marked in red to assess negatively correlated values to DV; if in black, the variables have a positive relationship.*

What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

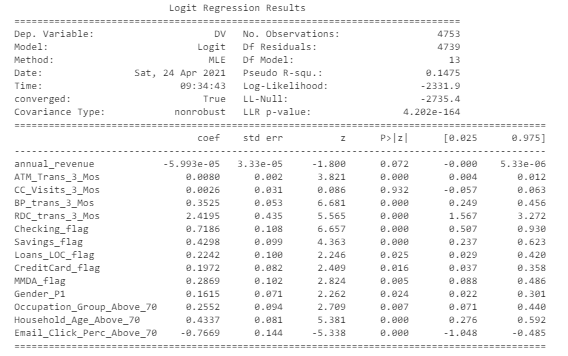
**Pseudo R-squared =** 0.1860

**Confusion Matrix:** 75.47% Accuracy

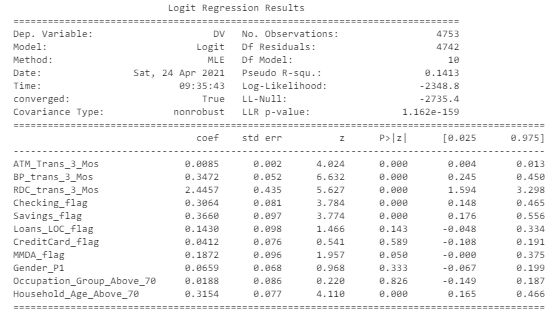


# Reiterating Diagnostic Plots

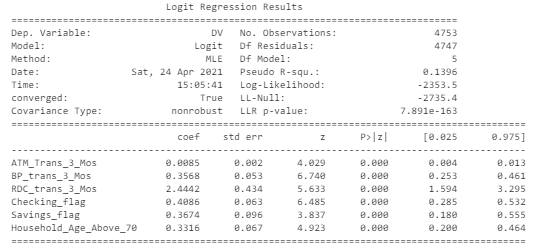
**Third Diagnostic Plot** – Removed insignificant variables and assessed the outcome.



**Forth Diagnostic Plot** – Removed insignificant variables and assessed the outcome.



**Fifth Diagnostic Plot** – Removed insignificant variables and assessed the outcome.

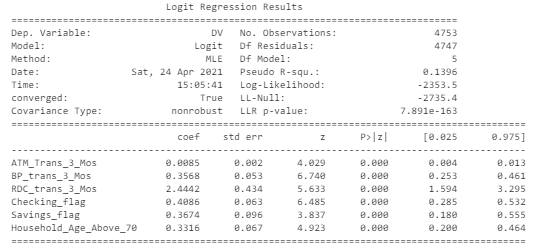


Conclusion: I will run the 5th diagnostic plot up against the first two models before concluding this report.

# Third Model Run

## **Diagnostic Plot**

Show diagnostic plots of your predictors



## **Significant and Non-Significant Variables**

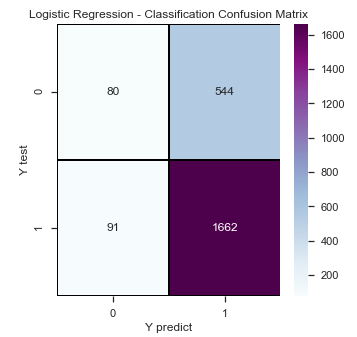
What are significant variables in the model in predicting customer value? Not significant?

|  |
| --- |
| Significant |
| ATM\_trans\_3\_Mos |
| BP\_trans\_3\_Mos |
| RDC\_trans\_3\_Mos |
| Checking\_flag |
| Savings\_flag |
| Household\_Age\_Above\_70 |

## **Variables Review**

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?

No, all variables are positive and there are no conflicting relationships with DV.

What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

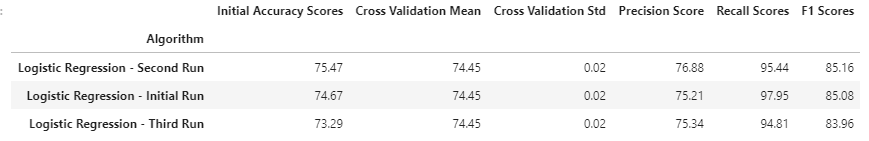
**Pseudo R-squared =** 0.1396

**Confusion Matrix:** 73.29% Accuracy 🡪

# Final Model Results – Interpretation

## **Improvement of model performance from your initial model**

First, I am going to assess the performance of each of the three models used:



Although the first two models had the best results, I am going to go with the third model’s results.

Why? Two reasons:

* The accuracy, precision, recall and F1 scores are all very similar, and
* With only 5 variables used on the third run, I am much more confident in the model being more fit for the analysis I am conducting. Essentially, I will would overfitting the model while achieving virtually the same results if I decided to use the first two models versus the third.

It is also worth noting cross validation did occur (n=10) and the mean values for accuracy were equivalent for the three models as well.

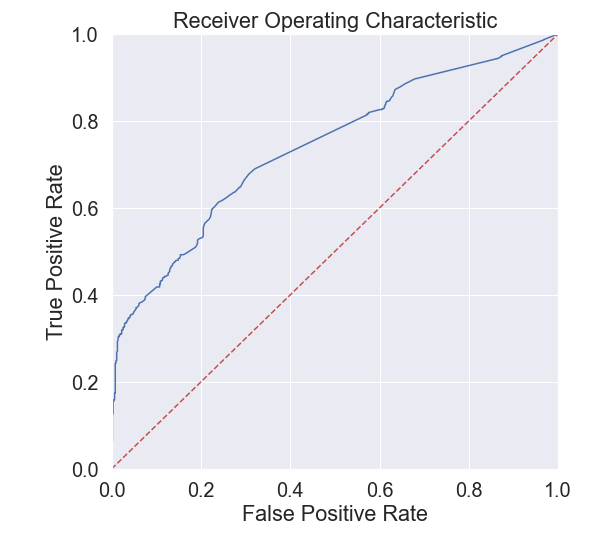
## **Interpretation of the model**

The accuracy score was 73.29, which means it was able to accurately predict the outcome (0 or 1) for if a customer did or did not have mobile banking.

The precision score was 75.34, which means it able to select the 3 out of 4 true positive predictions out of the predicted true and false positives. Essentially, 1 out of 4 were false positive predictions (a customer does have mobile or online banking when they do not) was a true positive (a customer does have mobile banking and should have been guessed as so).

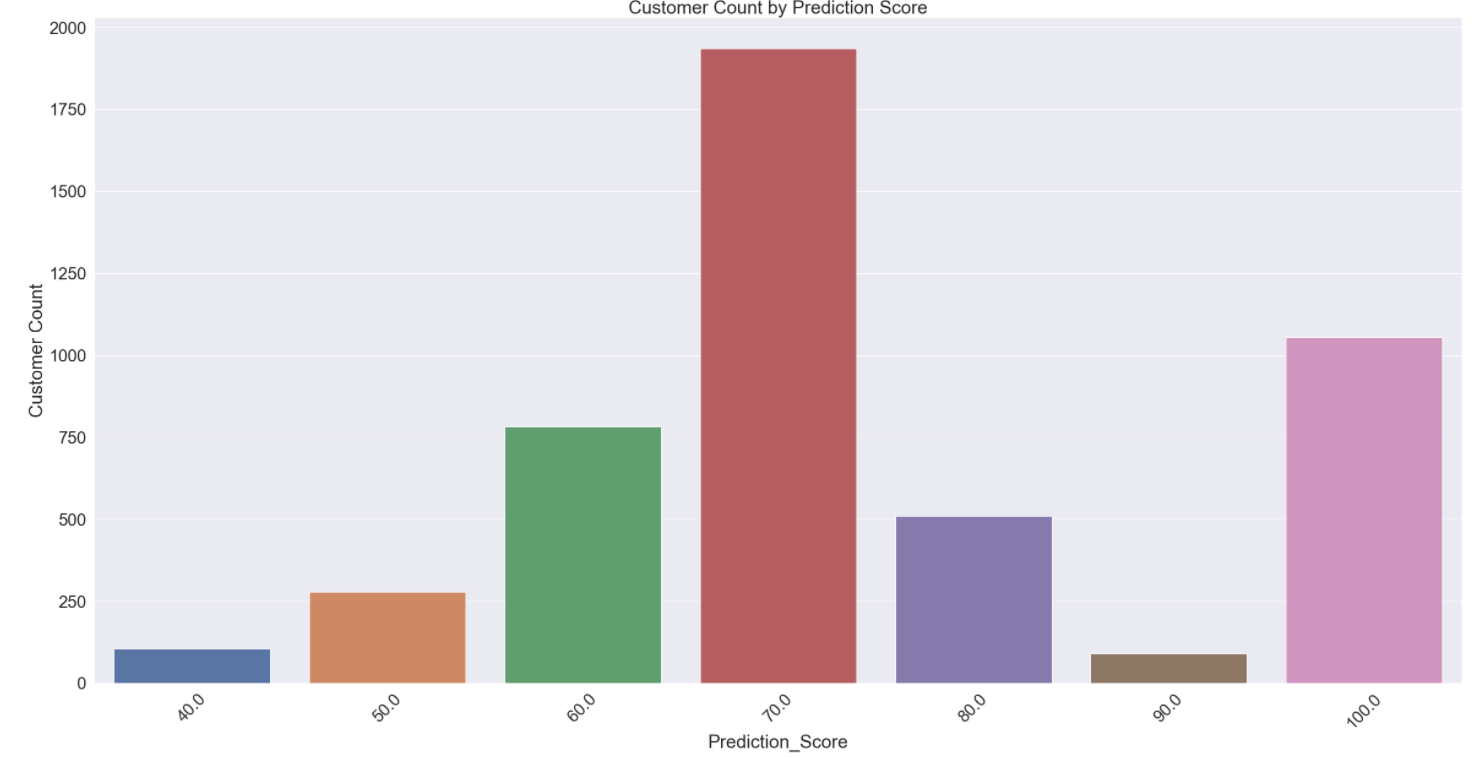
The recall score was 94.81, which means approximately 95 out of 100 relevant items were selected. Essentially, 5 out of 100 false negative predications (a customer does not have mobile or online banking when they do) was actually a true positive (a customer does have mobile banking and should have been guessed as so).

The ROC curve also indicates positive results.



## **Management Recommendations**

## How will you use the model to support the Bank’s digital channel migration campaign?



A customer has been assigned a prediction score, which is the probability a customer has mobile or online banking rounded to the nearest 10th. For example, there are just a little over 1,000 customers with a 100% chance of having mobile or online banking, while about 100 and 275 customers with a 40 and 50 percent chance of having mobile or online banking, respectively.

Below are the relevant featured used in this analysis for management to be able to strategize:

* **ATM\_Trans\_3\_Mos** – The number of completed ATM transactions in the last 3 months.
* **BP\_Trans\_3\_Mos** – The number of completed remote deposit check transactions in the last 3 months.
* **RDC\_Trans\_3\_Mos** – The number of completed remote deposit check transactions in the last 3 months.
* **Checking\_flag –** The customer has a checking account.
* **Savings\_flag** – The customer has a Savings account.
* **Housedhold\_Age\_Over\_70** – Individuals where the head of household has over 70% change of having DV; this includes households where the head of household is under the age of 65.

Essentially, it boils down to whether or not an individual under the age of 65 who has a checking and savings account, as well as the acumen to utilize ATMs, bill pay, and remote deposits that helps management predict the questions they are looking to answer.

If you are looking to get the best dollar spend, the two recommended strategies are to target:

* Younger individuals who do not use the above services and potentially provide training
* Assess if the online migration is going to be value added for older individuals who might have a lower tolerance for accepting training or change