**Matt Purvis – BUA 6110 – Predictive & Prescriptive Analytics – Final Project DDD and Results**

Table of Contents

[DDD Summary Steps 2](#_Toc70969852)

[Character Variables 2](#_Toc70969853)

[Missing Values 2](#_Toc70969854)

[Outliers 2](#_Toc70969855)

[Relationships 3](#_Toc70969856)

[Dataset 3](#_Toc70969857)

[Objectives 3](#_Toc70969858)

[Initial Model Run 4](#_Toc70969859)

[Correlations 4](#_Toc70969860)

[Diagnostic Plot 4](#_Toc70969861)

[Significant and Non-Significant Variables 4](#_Toc70969862)

[Second Model Run 5](#_Toc70969863)

[Correlations 5](#_Toc70969864)

[Diagnostic Plot 6](#_Toc70969865)

[Significant and Non-Significant Variables 6](#_Toc70969866)

[Reiterating Diagnostic Plots 7](#_Toc70969867)

[Third Model Run 8](#_Toc70969868)

[Diagnostic Plot 8](#_Toc70969869)

[Significant and Non-Significant Variables 8](#_Toc70969870)

[Variables Review 8](#_Toc70969871)

[Final Model Results – Interpretation 9](#_Toc70969872)

[Improvement of model performance from your initial model 9](#_Toc70969873)

[Interpretation of the model 9](#_Toc70969874)

[How will you use the model to support the client’s efforts to better understand if a customer will churn? 10](#_Toc70969875)

[Management Recommendations 11](#_Toc70969876)

[Appendix 12](#_Toc70969877)

[Univariate Analysis – Visuals – Binary Values 12](#_Toc70969878)

[Univariate Analysis – Visuals – Multi-Value 12](#_Toc70969879)

[Univariate Analysis – Visuals – Continuous Values 13](#_Toc70969880)

# 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 | |
| Gender | SeniorCitizen |
| Partner | Dependents |
| PaperlessBilling |  |

## **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.

There are no missing values provided with this dataset.

## **Outliers**

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

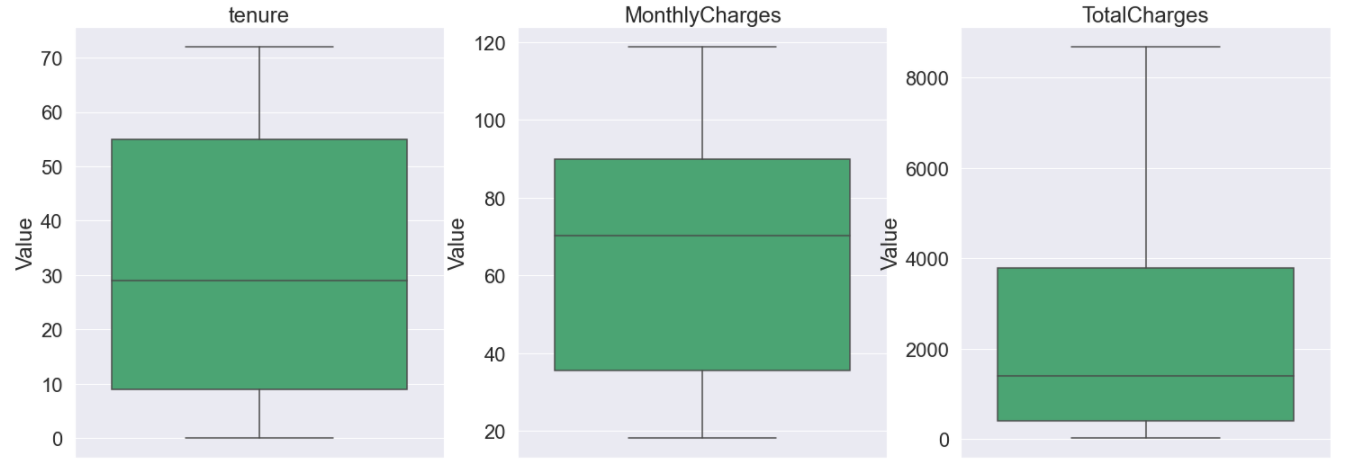
Outliers were assessed for three fields (tenure, MonthlyCharges, and TotalCharges). First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

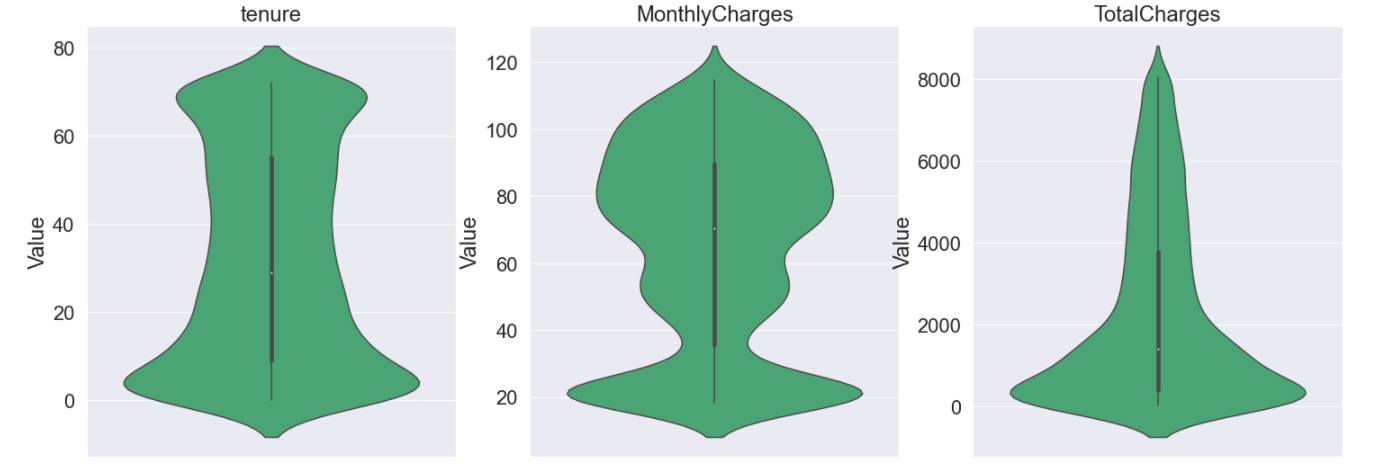
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in tenure is: 72.0
  + The 1st percentile in tenure is: 1.0
  + The 99th percentile in MonthlyCharges is: 114.7
  + The 1st percentile in MonthlyCharges is: 19.2
  + The 99th percentile in TotalCharges is: 8039.26
  + The 1st percentile in TotalCharges is: 19.9
* For each of the three variables, the following action was taken to remove any outliers.
  + Set a ceiling using the 99th percentile
  + Set a floor using 0 as the lowest allowable value

The below visuals 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 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

## **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.

# Objectives

**Objective**: Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs.

**Background**

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

# Initial Model Run

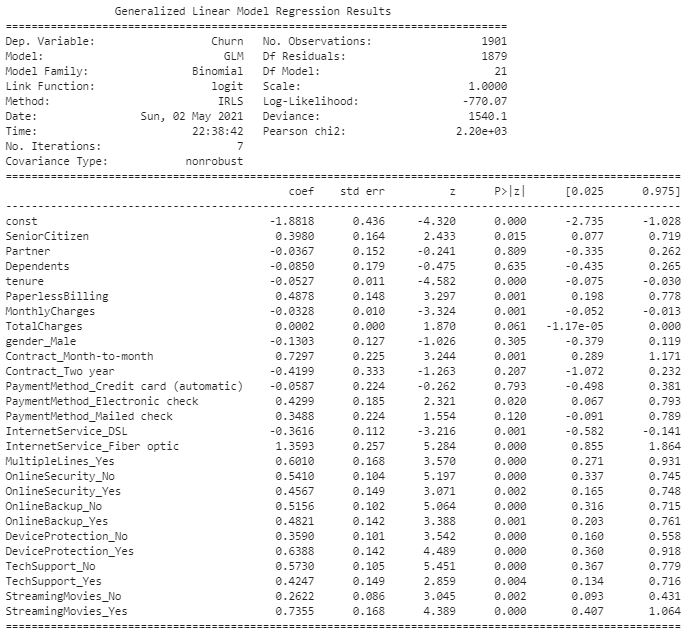
## **Correlations**

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

*See the ‘visuals folder’ that includes the correlation table used during the initial model run. Due to the size of the table, I have provided an PDF that is much easier to read than using a visual provided on this document.*

## **Diagnostic Plot**

Show diagnostic plots of your predictors

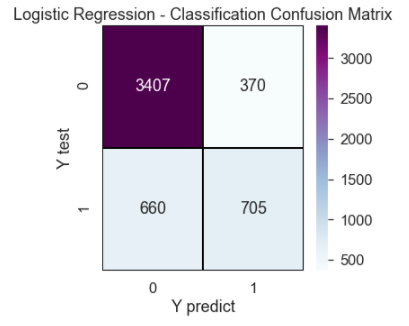
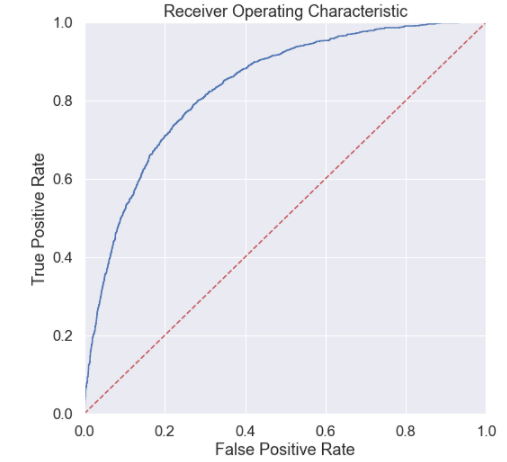


## **Significant and Non-Significant Variables**

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

|  |  |
| --- | --- |
| Significant | Non-Significant |
| SeniorCitizen | Partner |
| PaperlessBilling | Dependents |
| TotalCharges | Tenure |
| Contract\_Month-tomonth | MonthlyCharges |
| PaymentMethod\_Electronic check | Gender\_Male |
| InternetService\_Fiber optic | Contract\_Two year |
| MultipleLines\_Yes | PaymentMethod\_Credit card (automatic) |
| OnlineSecurity\_No | PaymentMethod\_Mailed check |
| OnlineSecurity\_Yes | InternetService\_DSL |
| OnlineBackup\_No | InternetService\_Fiber optic |
| OnlineBackup\_Yes | StreamingMovies\_No |
| DeviceProtection\_No |  |
| DeviceProtection\_Yes |  |
| TechSupport\_No |  |
| TechSupport\_Yes |  |
| StreamingMovies\_Yes |  |

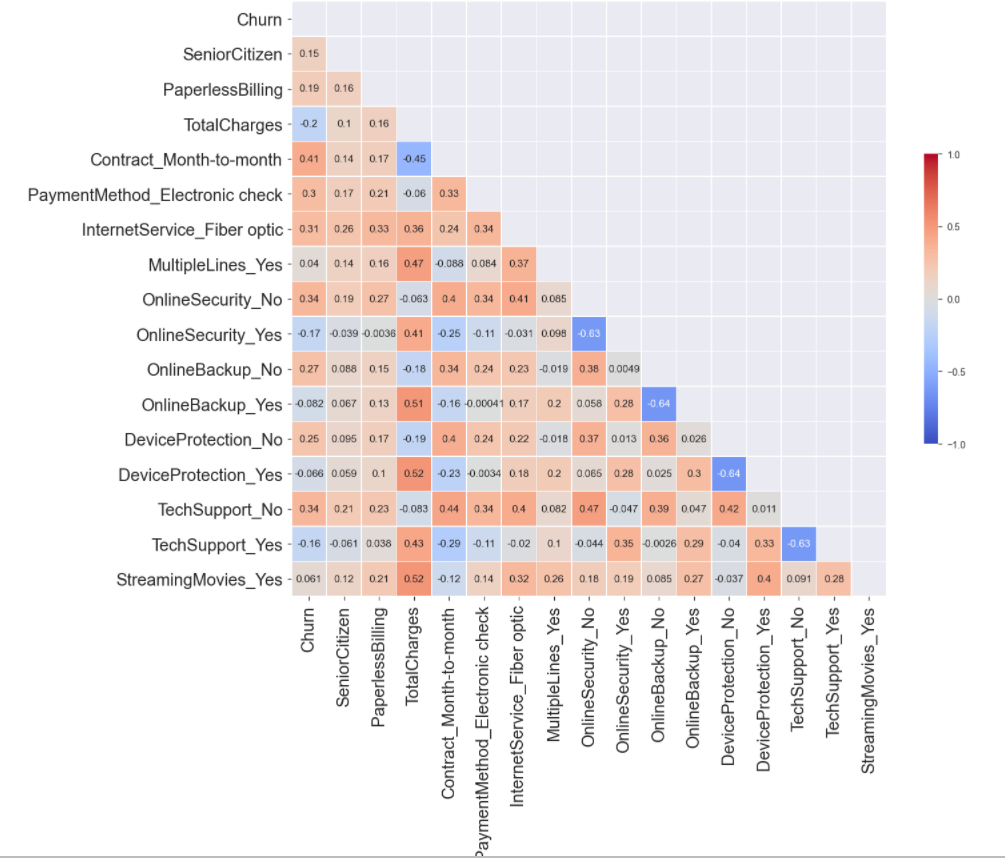
What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

# Second Model Run

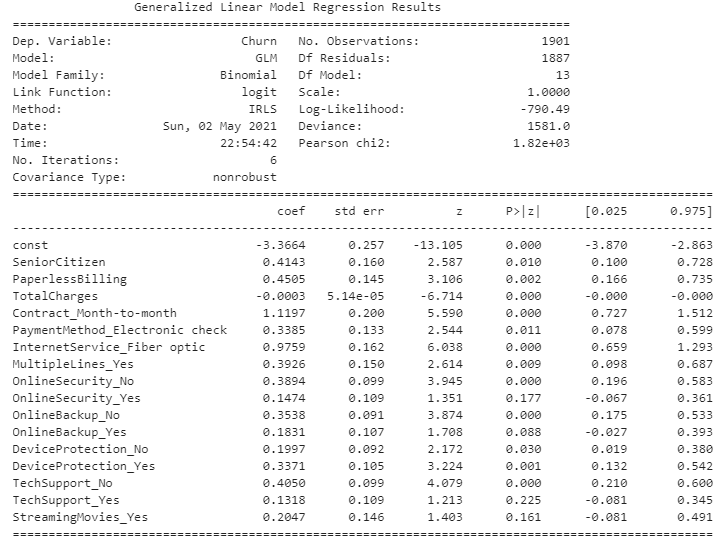
## **Correlations**

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



## **Diagnostic Plot**

Show diagnostic plots of your predictors

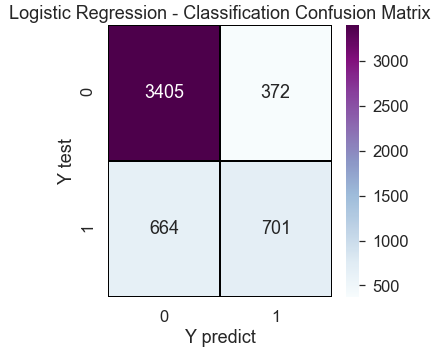
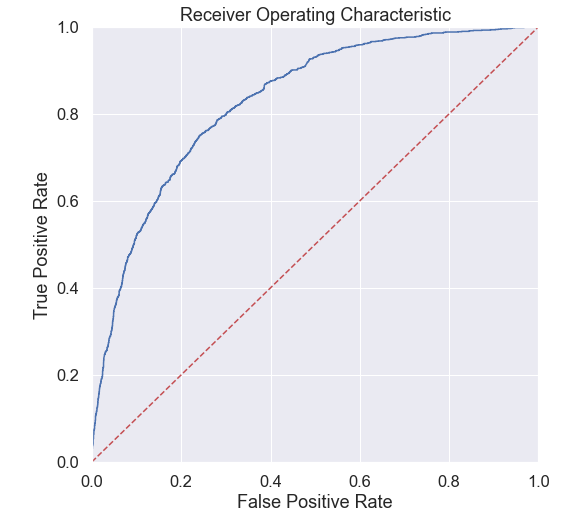


## **Significant and Non-Significant Variables**

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

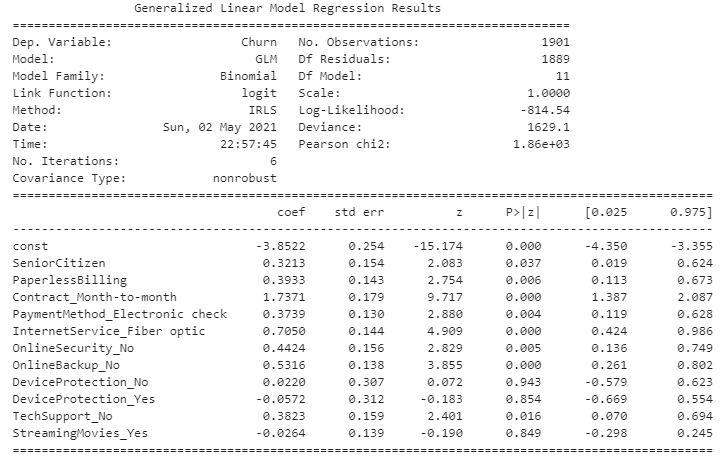
|  |  |
| --- | --- |
| Significant | Non-Significant |
| SeniorCitizen | **TotalCharges** |
| PaperlessBilling | **MultipleLines\_Yes** |
| Contract\_Month-tomonth | **OnlineSecurity\_Yes** |
| PaymentMethod\_Electronic check | **OnlineBackup\_Yes** |
| InternetService\_Fiber optic | **TechSupport\_Yes** |
| OnlineSecurity\_No |  |
| OnlineBackup\_No |  |
| DeviceProtection\_No |  |
| DeviceProtection\_Yes |  |
| TechSupport\_No |  |
| StreamingMovies\_Yes |  |

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

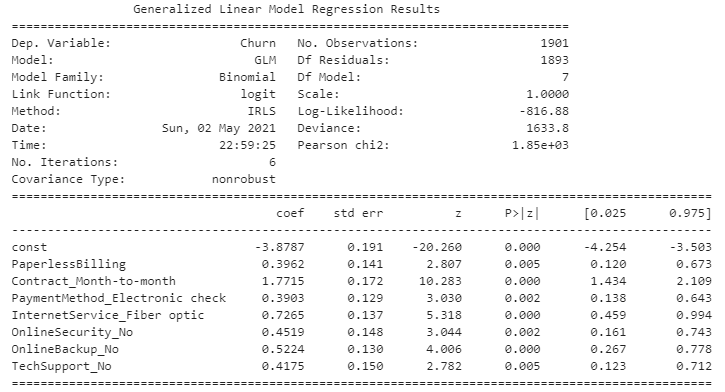
 

# Reiterating Diagnostic Plots

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



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

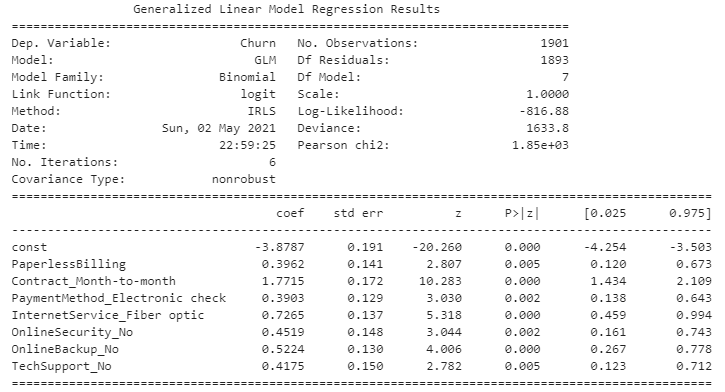


Conclusion: I will run the 4th diagnostic plot up against the first two models before concluding this report.

# Third Model Run

## **Diagnostic Plot**

Show diagnostic plots of your predictors (*4th model run used above*)



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

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

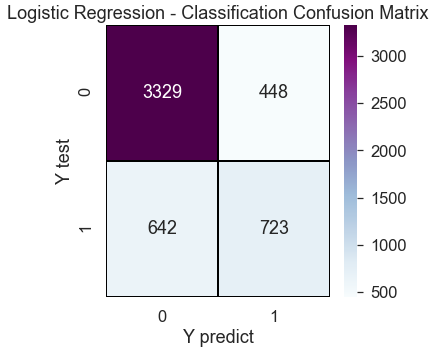
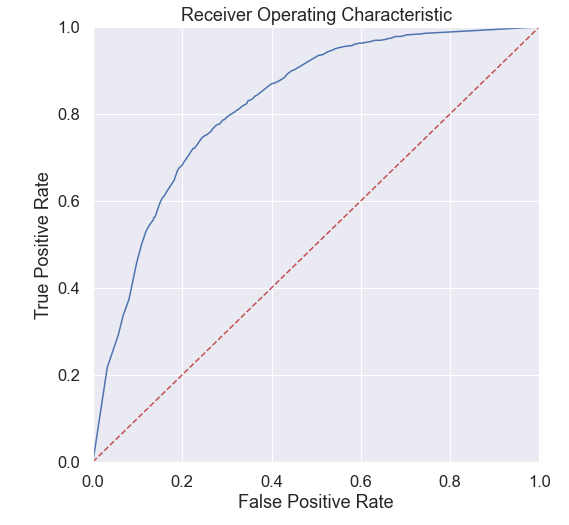
|  |
| --- |
| Significant |
| PaperlessBilling |
| Contract\_Month-to-month |
| PaymentMethod\_Electronic check |
| InternetService\_Fiber optic |
| OnlineSecurity\_No |
| OnlineBackup\_No |
| TechSupport\_No |

## **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.

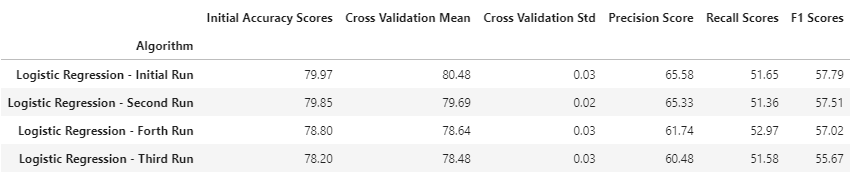
What is the goodness of fit (pseudo-R squared, ROC curve, confusion matrix) of the model?

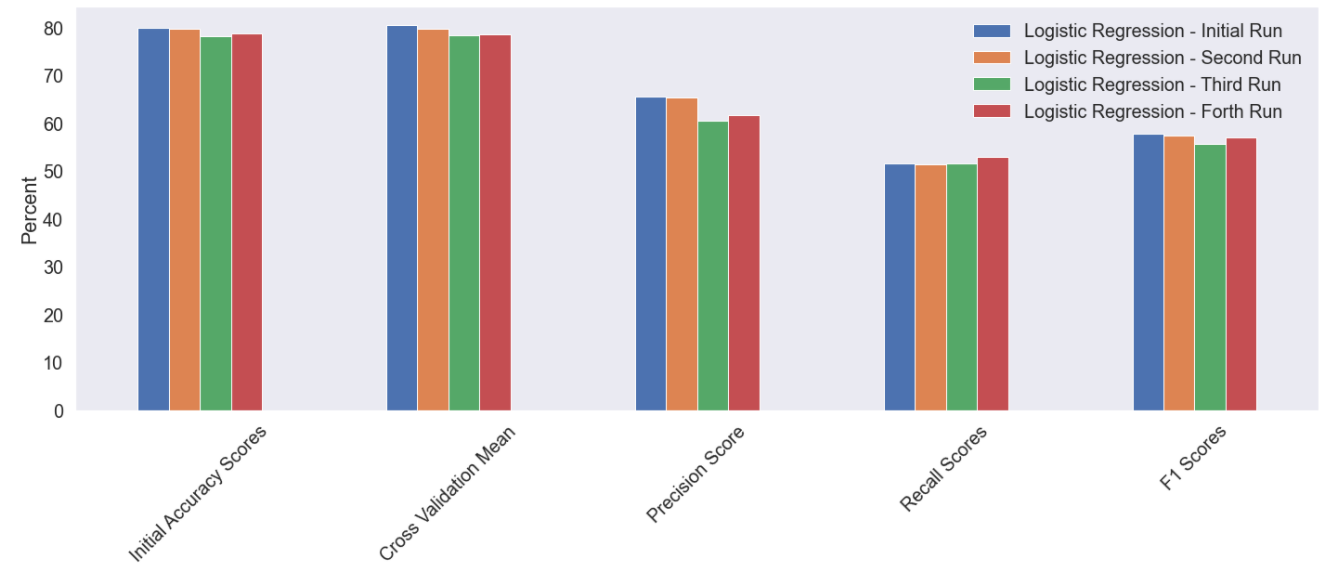
 

# Final Model Results – Interpretation

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

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





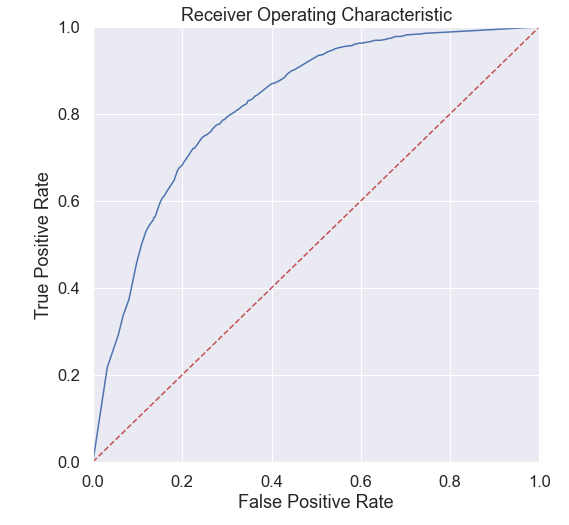
Although the four models had similar results, I am select to use the fourth model.

Why? Two reasons:

* The accuracy, precision, recall and F1 scores are all very similar, and
* With only 7 variables used on the fourth run, I am much more confident in the model being more fit for the analysis I am conducting. Essentially, I would be overfitting the model if I used any of the other models and achieving virtually the same accuracy score.

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

## **Interpretation of the model**

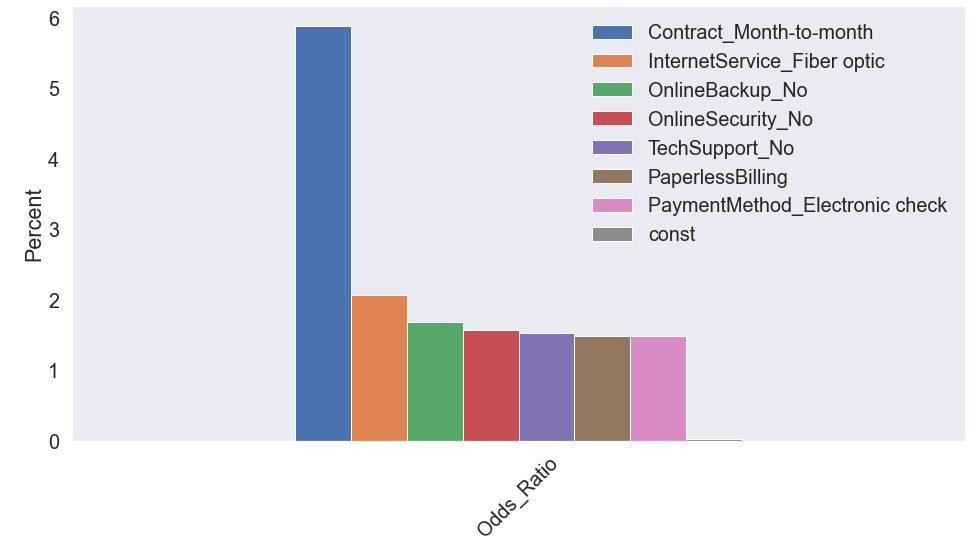
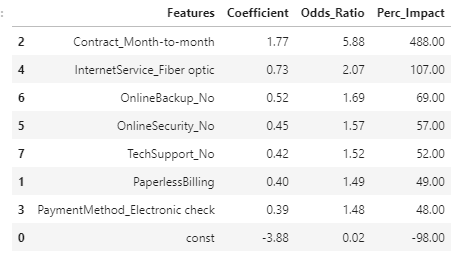
The accuracy score was 78.8, which means it was able to accurately predict the outcome (0 or 1) for if a customer did or did not churn.

The precision score was 61.74, which means it able to select the 6 out of 10 true positive predictions out of the predicted true and false positives. Essentially, 4 out of 10 were false positive predictions (a customer will churn when predicted they would not) was a true positive (a customer did churn and should have been guessed as so).

The recall score was 52.97, which means approximately 53 out of 100 relevant items were selected. Essentially, 47 out of 100 false negative predications (a customer did not churn when they actually do) was actually a true positive (a customer did churn and should have been guessed as so).

The ROC curve (right) also indicates positive results.

## How will you use the model to support the client’s efforts to better understand if a customer will churn?



Each factor has been assigned a score using the Coefficient and Odds Ratio to interpret the model’s results.

The features are sorted from most impactful to least impactful. Below is a summary of the variables for interpretation:

|  |  |
| --- | --- |
| Variables & Metrics | Interpretation |
| Contract\_Month-to month | The probability of a customer churning is 488% higher if a customer has a month-to-month contract versus the alternative (one year or two year), assuming all other variables constant. This is the most important variable in the analysis and will be communicated to the client. |
| Coefficient: 1.77 |
| Odds Ratio: 5.88 |
| InternetService\_Fiber optic | The probability of a customer churning is 107% higher if a customer has fiber optic service than not, assuming all over variables constant. This may be tied to the high-correlation between monthly charges and the service (*see appendix for univariate box plot of this variable*). |
| Coefficient: 0.73 |
| Odds Ratio: 2.07 |
| OnlineBackup\_No | The probability of a customer churning is 69% higher if a customer has no online backup, assuming all other variables constant. |
| Coefficient: 0.52 |
| Odds Ratio: 1.69 |
| OnlineSecurity\_No | The probability of a customer churning is 57% higher if a customer has no online security services, assuming all other variables constant. |
| Coefficient: 0.45 |
| Odds Ratio: 1.57 |
| TechSupport\_No | The probability of a customer churning is 49% higher if a customer has no tech support services, assuming all other variables constant. |
| Coefficient: 0.42 |
| Odds Ratio: 1.49 |
| PaperlessBilling | The probability of a customer churning is 49% higher if a customer has paperless billing, rather than not, assuming all other variables constant. |
| Coefficient: 0.40 |
| Odds Ratio: 1.49 |
| PaymentMethod\_Electronic check | The probability of a customer churning is 48% higher if a customer chooses to elect for the use of electronic check versus other payment methods, assuming all other variables constant. |
| Coefficient: 0.39 |
| Odds Ratio: 1.48 |

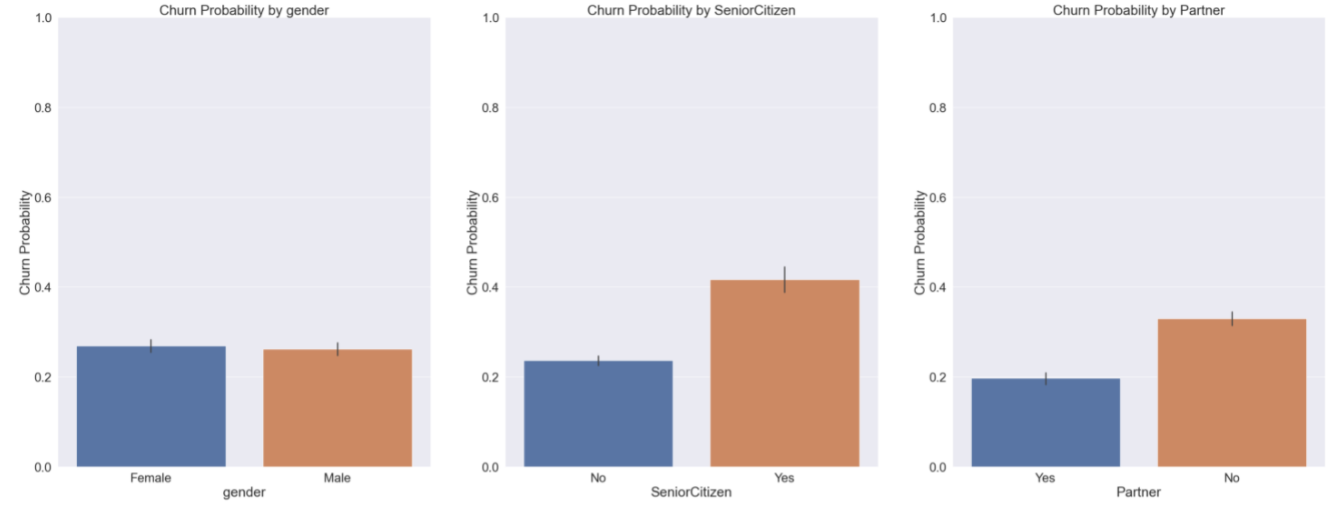
## **Management Recommendations**

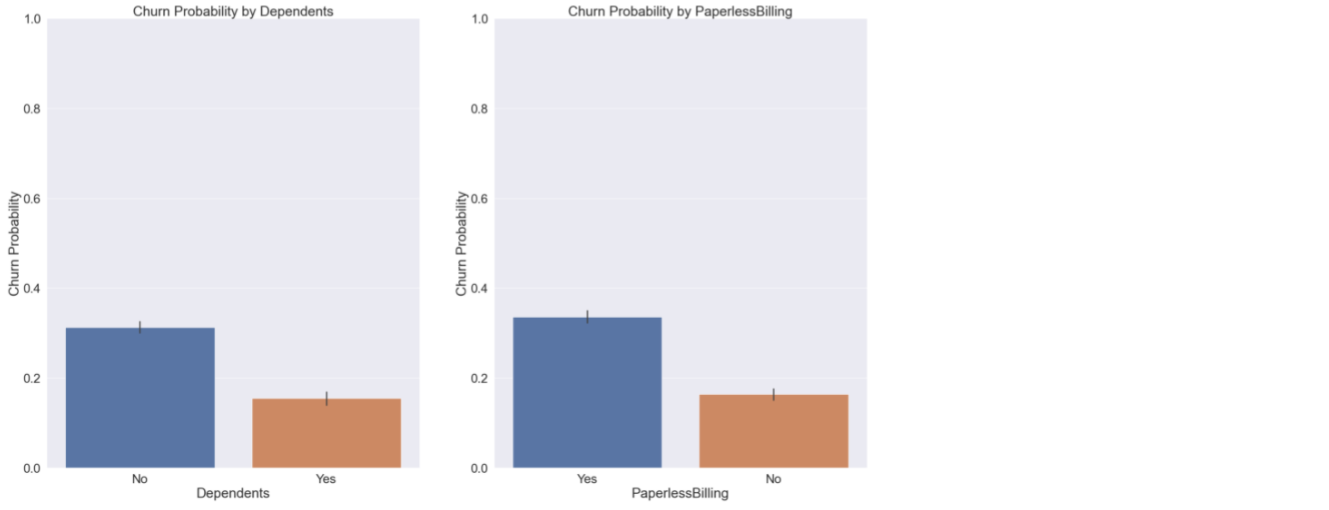
Although the model is subject to some degree of error, we are able to provide some recommendations to the client based on the results provided. Below are the recommendations to consider:

* First, continue to invest in further tuning the model by adding more historical data and using more feature engineering options that may be available. This will make the model, not only more accurate, but more reliable in the decision-making process, which can result in an even more thorough understanding of why a customer does or does not churn.
* Second, there is a much higher likelihood of a customer churning if they use the month-to-month service, so targeting these customers for longer-term contracts is highly recommended.
* Third, continue to research the relationship between high cost and fiber optics. Is there a way to price in a manner that might be a better value for customers, or is the fiber optic service just not providing a great experience for the customer? It could be a combination of both, so this is highly recommended to investigate.
* Forth, I am combining this recommendation for three of the variables listed above, but it is worth noting that they are not correlated. The solution just may be scalable for each. It is recommended to understand how you can provide these services to customers, as if a customer does not have online security, online backup, or tech support services, they are most likely to churn. They either give a customer peace of mind or are just a great value that maintains high customer satisfaction and results in less churning.
* Fifth, since a customer is more likely to choose a more “technology-friendly” approach by using paperless billing, it is recommended to research the process a customer goes through when electing this option. Why? If a customer is experiencing difficulties paying you, they are likely to de-value you and leave for a competitor.
* Finally, since a customer seems to churn by using an electronic check payment method versus any other, it is recommended, like the recommendation above, to, first, evaluate the customer experience using this method. If the customer experience is not poor, it might be prudent to offer some incentives to use another payment method.

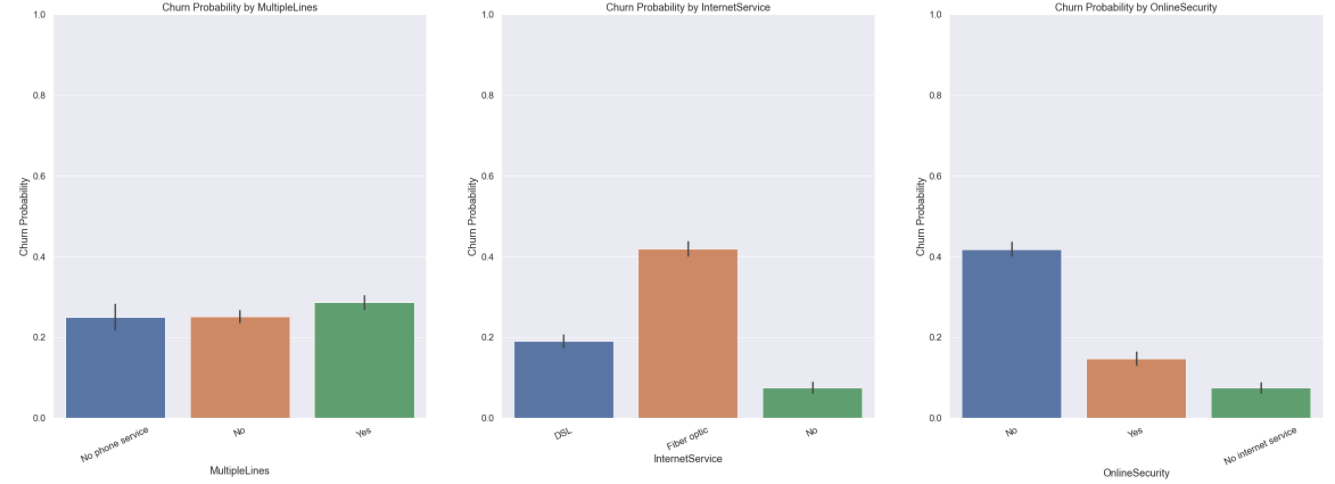
# Appendix

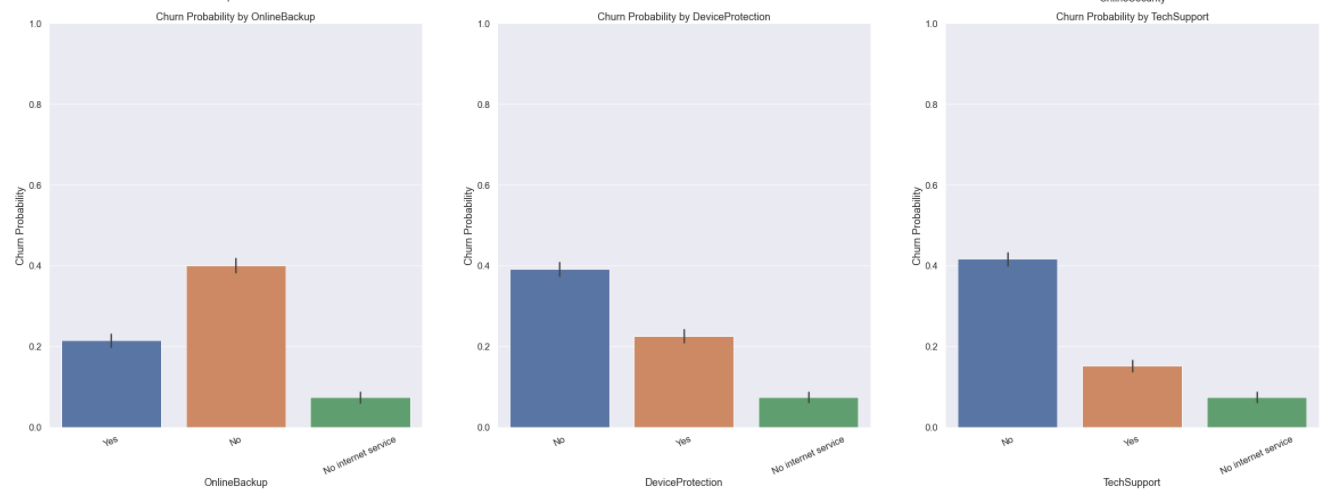
## **Univariate Analysis – Visuals – Binary Values**

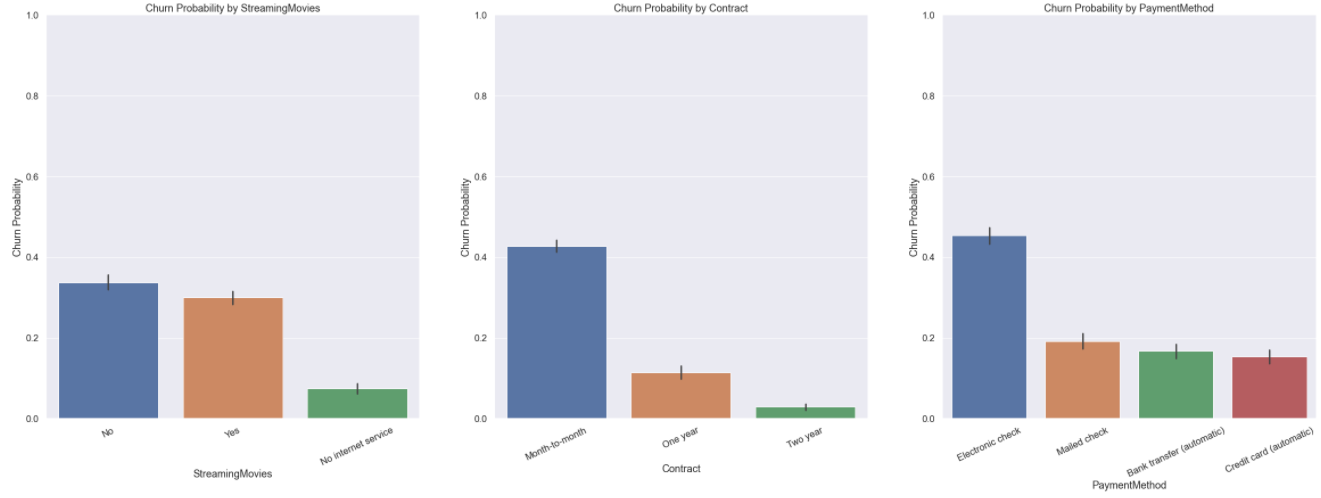




## **Univariate Analysis – Visuals – Multi-Value**







## **Univariate Analysis – Visuals – Continuous Values**

