**Mobility-Spending Classification**

Project Report

*Group 10*

*Student 1:* Parikshith T (91)-9901665577

thriyambaka.p@northeastern.edu

*Student 2:* Ekshitha Golla (91)-9553029192

golla.e@northeastern.edu

**Percentage of Effort Contributed by Student 1: 50% Percentage of Effort Contributed by Student 2: 50%**

Signature of Student 1: Parikshith T

Signature of Student 2: Ekshitha Golla

**Submission Date**: August 10, 2020

# Problem Setting:

The problem that was selected for analysis involves predicting the consumer spending behavior based on human mobility measures.

During initial stages of COVID-19, a greater number of people are tested as COVID positive. The main reason for this is not knowing how coronavirus is transmitted. After finding that coronavirus spreads between people through direct, indirect (through contaminated objects or surfaces), or close contact with infected people via mouth and nose secretions, government suggested people maintain at least 1metre distance from each other and to self – quarantine. This impacted the consumer spending behavior as most of the work places and shops are closed. The seven models, if successfully created and implemented, could potentially help in understanding the pattern of spending in this pandemic.

# Problem Definition:

In this project, the problem is to find how mobility measures impact consumer spending by building models, one for each of the response variables: spend\_acf, spend\_aer, spend\_apg, spend\_tws, spend\_all\_inchigh, spend\_all\_incmiddle, spend\_all\_inclow using mobility attributes as input variables. The main objective of the project is analyze and compare the performance of seven models.

# Data Sources:

The mobility variables indicate how visits to places, such as grocery stores and parks, are changing in each geographic region are provided by Google -<https://www.google.com/covid19/mobility/>.

The consumer spending measures are provided by affinity solutions and are used in opportunity insights database <https://github.com/OpportunityInsights/EconomicTracker>. For this project use the encoded dataset mobility-spending-encoded.csv. The encoded dataset has 1s and 0s introduced in the variables related to consumer spending. All values less than -0.1 are replaced by 0s. Similarly, all values greater than -0.1 are replaced by 1s. A 0 indicates a large drop in consumer spending and a 1 indicates a small drop in consumer spending.

# Data Description:

The following table details the attributes that were present in the original dataset.

*Table 1:* ***Mobility measures***

|  |  |
| --- | --- |
| **Attribute** | **Definition** |
| gps\_away\_from\_home | Time spent outside of residential locations. |
| gps\_retail\_and\_recreation | Time spent at retail and recreation locations. |
| gps\_grocery\_and\_pharmacy | Time spent at grocery and pharmacy locations. |
| gps\_parks | Time spent at parks. |
| gps\_transit\_stations | Time at inside transit stations. |
| gps\_workplaces | Time spent at work places. |
| gps\_residential | Time spent at residential locations. |

*Table 2:* ***Consumer Spending Measures***

|  |  |
| --- | --- |
| **Attribute** | **Definition** |
| spend\_all | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in all merchant category codes (MCC), 7 day moving average. |
| spend\_acf | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in accomodation and food service (ACF) MCCs, 7 day moving average, 7 day moving average. |
| spend\_aer | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in arts, entertainment, and recreation (AER) MCCs, 7 day moving average. |
| spend\_apg | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in general merchandise stores (GEN) and apparel and accessories (AAP) MCCs, 7 day moving average. |
| spend\_grf | Time at inside transit Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in grocery and food store (GRF) MCCs, 7 day moving average. |
| spend\_hcs | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in health care and social assistance (HCS) MCCs, 7 day moving average. |
| spend\_tws | Seasonally adjusted credit/debit card spending relative to January 4-31 2020 in transportation and warehousing (TWS) MCCs, 7 day moving average. |
| spend\_all\_inchigh | Seasonally adjusted credit/debit card spending by consumers living in ZIP codes with high (top quartile) median income, relative to January 4-31 2020 in all merchant category codes (MCC), 7 day moving average. |
| spend\_all\_incmiddle | Seasonally adjusted credit/debit card spending by consumers living in ZIP codes with middle (middle two quartiles) median income, relative to January 4-31 2020 in all merchant category codes (MCC), 7 day moving average. |
| spend\_all\_inclow | Seasonally adjusted credit/debit card spending by consumers living in ZIP codes with low (bottom quartiles) median income, relative to January 4-31 2020 in all merchant category codes (MCC), 7 day moving average |

# Data Exploration and Processing:

The first step of Data exploration is to understand the data, the type of input variables and output variables. In this project the input variables are mobility measures which are continuous variables and output variable or response variable is consumer spending variable which is a categorical variable.

As part of data preprocessing missing values should be handled, either by deleting rows with missing values or by assigning median as value to continuous variable and mode as value to categorical variable.

*Table 3: Variable and number of missing records*

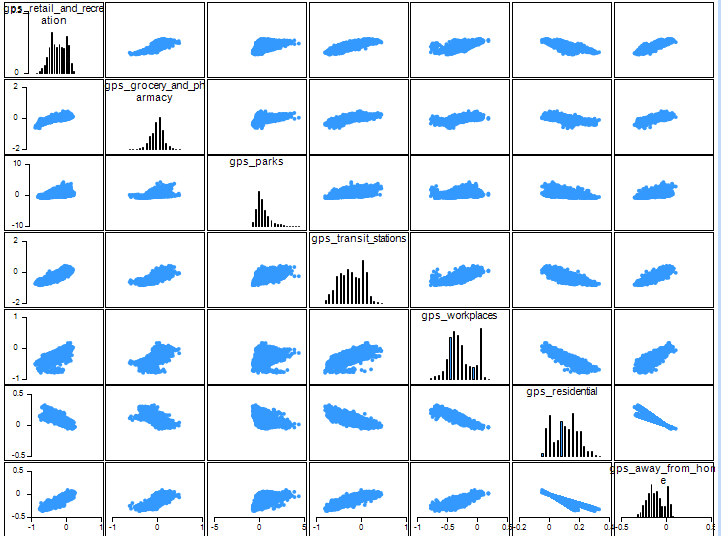
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | gps\_parks | spend\_all\_inchigh | spend\_all\_inclow |
| **# Missing Records** | 25 | 131 | 262 |

Data set has total of 6681 records, out of which there 418 records are having missing values. Using XLMiner, transform feature, we can delete these records.

*Table 4: Handling missing records*

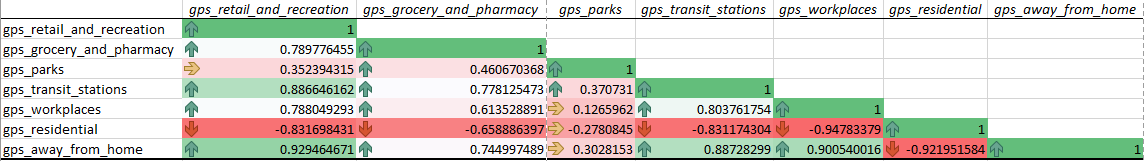
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | gps\_parks | spend\_all\_inchigh | spend\_all\_inclow | Other variables |
| **Reduction Type** | DELETE RECORD | DELETE RECORD | DELETE RECORD | NONE |
| **# Records Treated** | 25 | 131 | 262 | 0 |
| **Missing Value Code** | NA | | |  |
| **# Output Records** | 6263 | | |  |
| **#Records Deleted** | 418 | | |  |

After handling missing values, the first visualization tool implemented was a scatterplot matrix.



*Figure 1: Scatterplot Matrix evaluating attribute correlation*

Correlation matrix:



Based on the above scatter plot and correlation matrix, we can observe that gps\_away\_from\_home, gps\_transit\_stations and gps\_retail\_and\_recreation is strongly correlated.

The data exploration and visualization tools utilized in the initial data examination assist in the eventual determination of which attributes to focus on as predictors when building the classification model.

# Data Mining Tasks:

# As empty records are handled as part of data preprocessing, the next step is to partition data.

# For all the models, data is partitioned into 3 parts – 70% training data, 15% validation data and 15% testing data.

# Model 1 (response variable - spend\_acf)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_acf as output variable. Here, (based on feature importance) we can ignore gps\_retail\_and\_recreation, gps\_transit\_stations as they as highly correlated with gps\_away\_from\_home.

*Table 5: Variables used for building Classification tree for response variable spend\_acf*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential | gps\_away\_from\_home |
| **Output Variable** | spend\_acf | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_grocery\_and\_pharmacy” and “gps\_parks”

*Table 6: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_grocery\_and\_pharmacy** | 0.179744526 |
| **gps\_parks** | 0.263001825 |
| **gps\_workplaces** | 0.333257299 |
| **gps\_residential** | 0.036268248 |
| **gps\_away\_from\_home** | 0.218065693 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure 2: Fully grown tree*

Based on the above tree, we can create rules to predict if new record results in large drop or small drop in accommodation and food service spending.

# Model 2 (response variable - spend\_aer)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_aer as output variable. Here, we can ignore gps\_away\_from\_home, gps\_transit\_stations as they as highly correlated with gps\_retail\_and\_recreation.

*Table 7: Variables used for building Classification tree for response variable spend\_aer*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_retail\_and\_recreation | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential |
| **Output Variable** | spend\_aer | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_retail\_and\_recreation” and “gps\_grocery\_and\_pharmacy”

*Table 8: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_retail\_and\_recreation** | 0.993385036 |
| **gps\_grocery\_and\_pharmacy** | 0.981751825 |
| **gps\_parks** | 0.433166058 |
| **gps\_workplaces** | 0.563640511 |
| **gps\_residential** | 0.342381387 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

# Model 3 (response variable - spend\_apg)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_aer as output variable. Here, we can ignore gps\_away\_from\_home, gps\_transit\_stations as they as highly correlated with gps\_retail\_and\_recreation.

*Table 9: Variables used for building Classification tree for response variable spend\_apg*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_retail\_and\_recreation | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential |
| **Output Variable** | spend\_apg | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_retail\_and\_recreation” and “gps\_grocery\_and\_pharmacy”

*Table 10: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_retail\_and\_recreation** | 1.244525547 |
| **gps\_grocery\_and\_pharmacy** | 1.093978102 |
| **gps\_workplaces** | 1.042427007 |
| **gps\_parks** | 0.892791971 |
| **gps\_residential** | 0.480839416 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

# Model 4 (response variable - spend\_tws)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_aer as output variable. Here, we can ignore, gps\_retail\_and\_recreation, gps\_transit\_stations as they as highly correlated with gps\_away\_from\_home.

*Table 11: Variables used for building Classification tree for response variable spend\_tws*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential | gps\_away\_from\_home |
| **Output Variable** | spend\_tws | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_parks” and “gps\_away\_from\_home”

*Table 12: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_parks** | 0.960994526 |
| **gps\_away\_from\_home** | 0.826870438 |
| **gps\_workplaces** | 0.510036496 |
| **gps\_grocery\_and\_pharmacy** | 0.419251825 |
| **gps\_residential** | 0.057481752 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

# Model 5 (response variable - spend\_all\_inchigh)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_all\_inchigh as output variable. Here, we can ignore, gps\_away\_from\_home, gps\_transit\_stations as they as highly correlated with gps\_retail\_and\_recreation.

*Table 13: Variables used for building Classification tree for response variable spend\_all\_inchigh*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_retail\_and\_recreation | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential |
| **Output Variable** | spend\_all\_inchigh | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_workplaces” and “gps\_paks”

*Table 14: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_workplaces** | 1.113366788 |
| **gps\_parks** | 0.900547445 |
| **gps\_retail\_and\_recreation** | 0.807709854 |
| **gps\_grocery\_and\_pharmacy** | 0.658759124 |
| **gps\_residential** | 0.438868613 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

# Model 6 (response variable - spend\_all\_incmiddle)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_all\_incmiddle as output variable. Here, we can ignore, gps\_away\_from\_home, gps\_transit\_stations as they as highly correlated with gps\_retail\_and\_recreation.

*Table 15: Variables used for building Classification tree for response variable spend\_all\_incmiddle*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_retail\_and\_recreation | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential |
| **Output Variable** | spend\_all\_incmiddle | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_grocery\_and\_pharmacy” and “gps\_retail\_and\_recreation”

*Table 16: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_grocery\_and\_pharmacy** | 1.484489051 |
| **gps\_retail\_and\_recreation** | 1.45415146 |
| **gps\_workplaces** | 1.363138686 |
| **gps\_parks** | 1.052919708 |
| **gps\_residential** | 0.936359489 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

# Model 7 (response variable - spend\_all\_inclow)

# Using Data mining, classify feature, we can build classification tree considering five mobility variables as input variables and spend\_all\_incmiddle as output variable. Here, we can ignore, gps\_away\_from\_home, gps\_transit\_stations as they as highly correlated with gps\_retail\_and\_recreation.

*Table 17: Variables used for building Classification tree for response variable spend\_inclow*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | | | | | |
| **# Variables** | 5 | | | | |
| **Scale Variables** | gps\_retail\_and\_recreation | gps\_grocery\_and\_pharmacy | gps\_parks | gps\_workplaces | gps\_residential |
| **Output Variable** | spend\_all\_inclow | | | | |

# Of the set of variables included, each was given a relative importance in the model. The most important variables turned out to be “gps\_workplaces” and “gps\_retail\_and\_recreation”

*Table 18: Classification tree predicting feature importance*

|  |  |
| --- | --- |
| **Feature** | **Importance** |
| **gps\_workplaces** | 2.180885036 |
| **gps\_retail\_and\_recreation** | 1.881386861 |
| **gps\_parks** | 1.397810219 |
| **gps\_grocery\_and\_pharmacy** | 1.167427007 |
| **gps\_residential** | 0.784899635 |

# Fully grown tree was created as part of this analysis, displaying decision node values and final output variable terminals.

# 

*Figure: Fully grown tree*

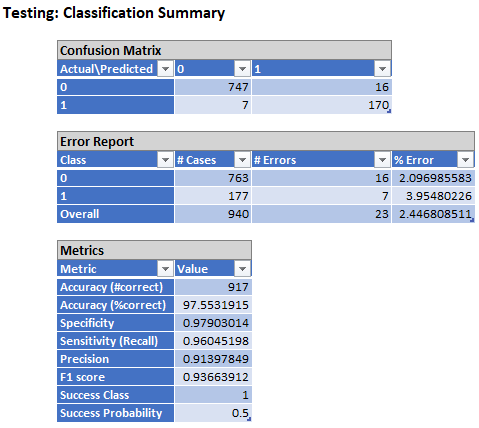
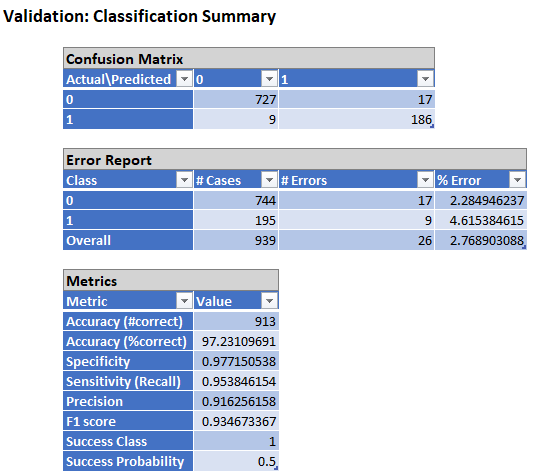
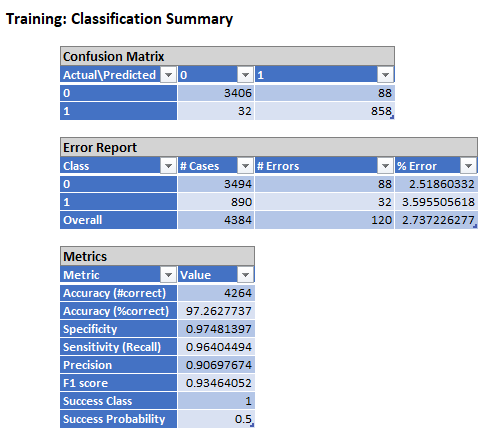
**Performance Evaluation:**

The following sections report the performance metrics found for each of the models built.

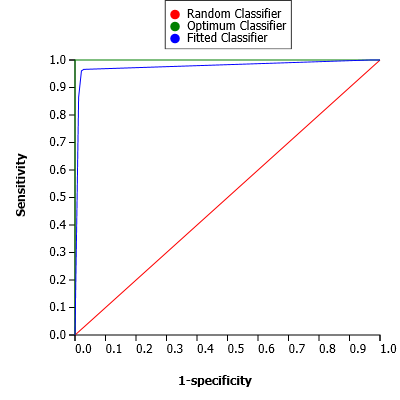
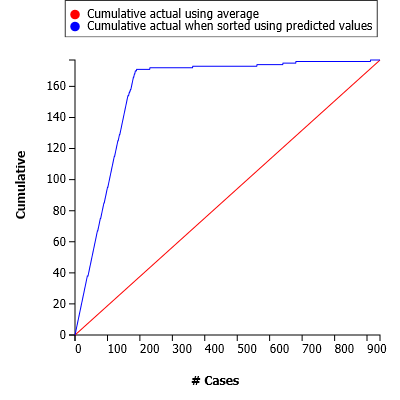
# Model 1 (response variable - spend\_acf)

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

*Table 19: Confusion matrix, error report and metrics for Model 1*



The model shows 2% error rate in all the data sets. The model is performing good on training, validation and testing sets with an accuracy of 97%.



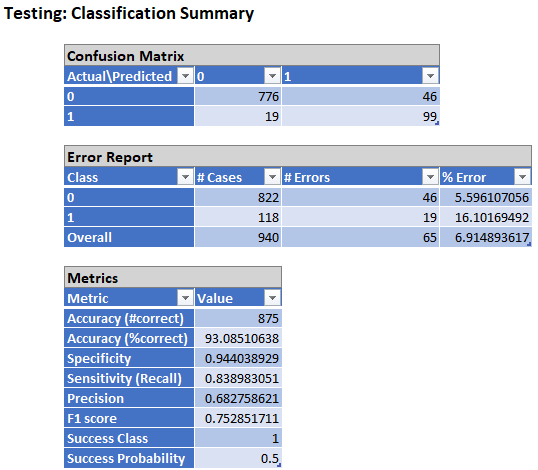
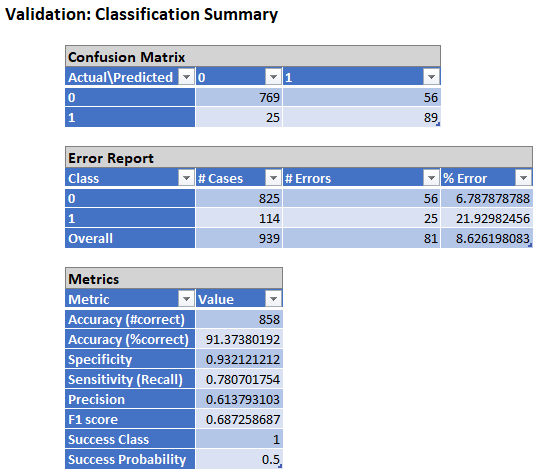
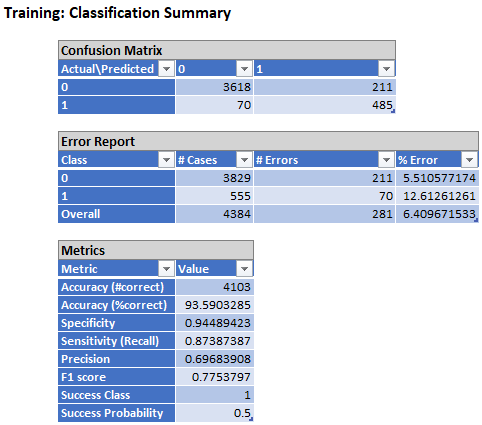
*Figure: Lift chart and ROC curve for model with AUROC = 0.976*

The AUROC for the model is 0.97, which shows the percentage of correctly classified records. As another indicator of model performance, the lift chart compares the number of correctly classified cases to the total number of records. The higher the lift, the better the model’s performance. In this case, the lift chart is far from the red line, which indicates that the model is performing good.

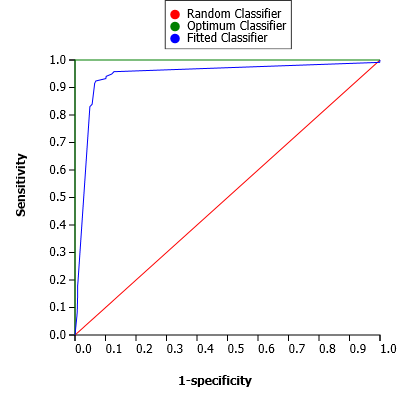
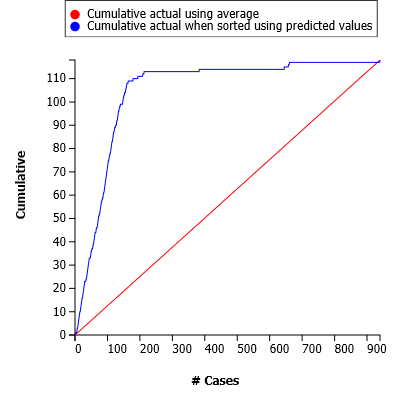
# Model 2 (response variable - spend\_aer)

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

*Table 20: Confusion matrix, error report and metrics for Model 2*



The model shows 6% error rate in training and testing data sets. But error rate increased to 8% in validation data set, this might be due to overfitting of training data. The model is performing better on training and testing sets with an accuracy of 93%.



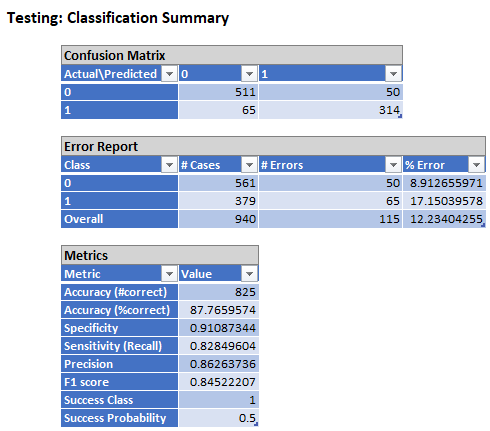
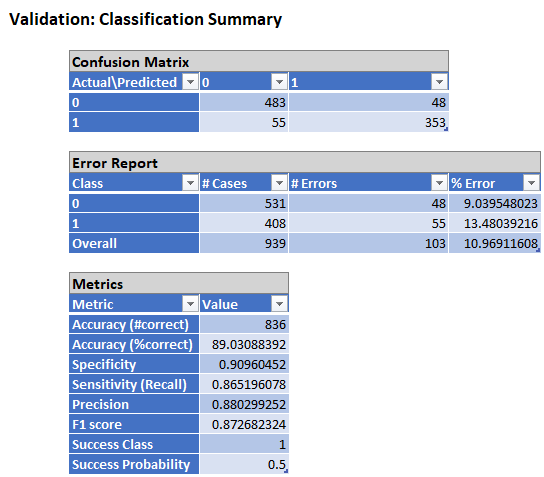
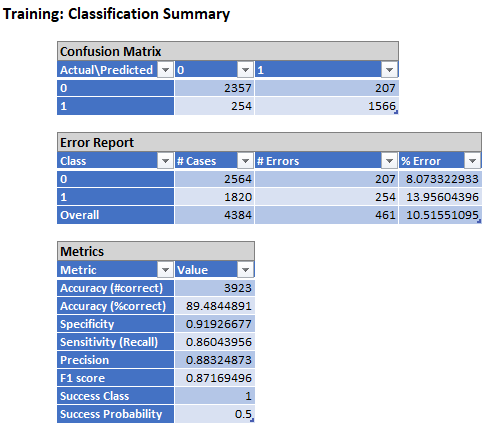
*Figure: Lift chart and ROC curve for model with AUROC = 0.9441*

The AUROC for the model is 0.9334, which shows the percentage of correctly classified records. Lift chart is far from the red line, which indicates that the model is performing better.

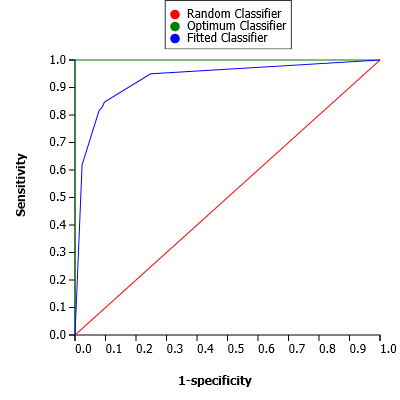
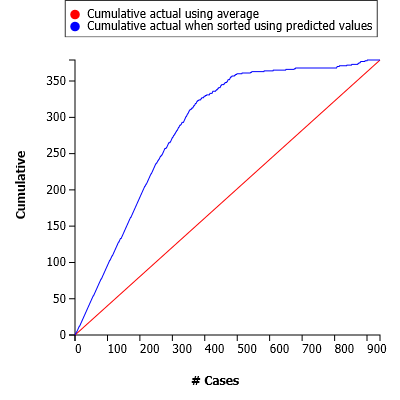
**Model 3 (response variable - spend\_apg)**

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

*Table 21: Confusion matrix, error report and metrics for Model 3*



The model shows 10% error rate in training and validation data sets. But error rate increased to 12% in testing data set, this might be due to overfitting of data. The model is performing better on training and testing sets with an accuracy of 87%.



*Figure: Lift chart and ROC curve for model with AUROC = 0.9311*

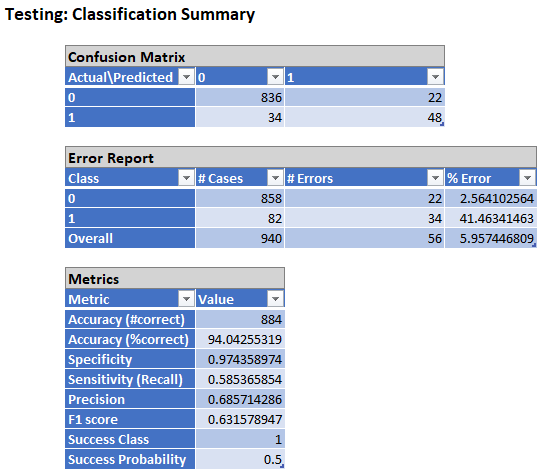
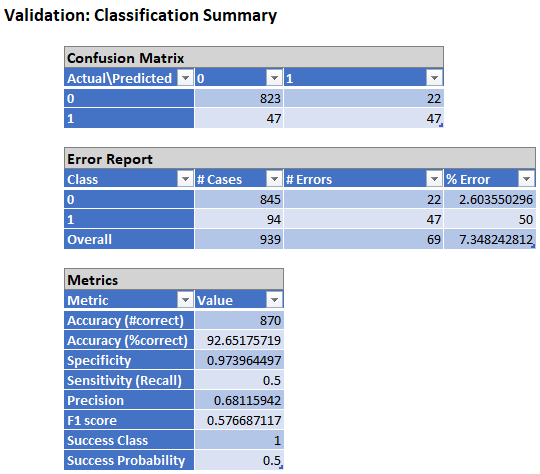
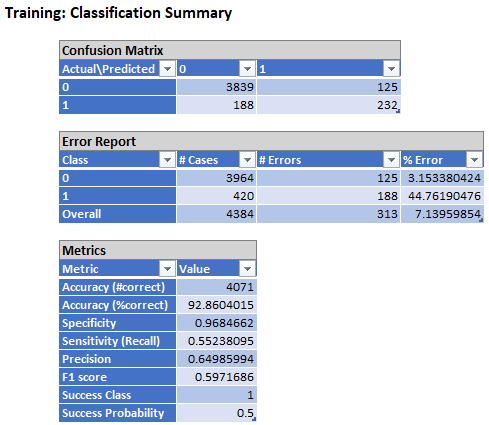
The AUROC for the model is 0.8811, which shows the percentage of correctly classified records. Lift chart is not much lifted from the red line, which indicates that the model is performing average.

# Model 4 (response variable - spend\_tws)

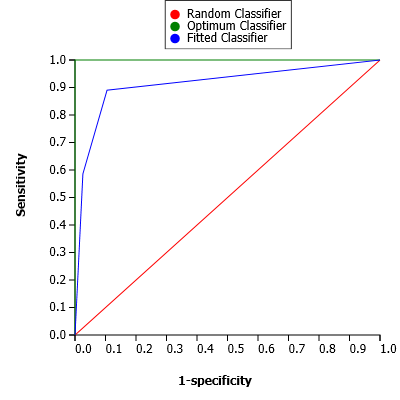
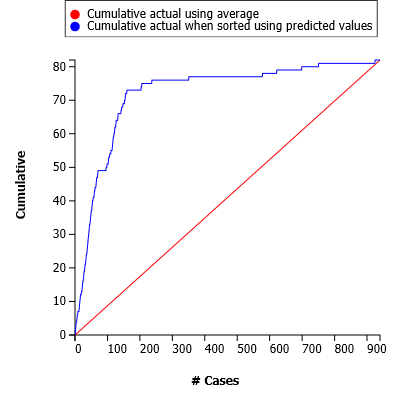
The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

The model shows 7% error rate in training and validation data sets. But error rate decreased to 6% in validation data set. The model is performing good on testing sets with an accuracy of 94%.

*Table 22: Confusion matrix, error report and metrics for Model 4*



The AUROC for the model is 0.9111, which shows the percentage of correctly classified records. Lift chart is not much more lifted than the red line representing the cumulative actual using average.



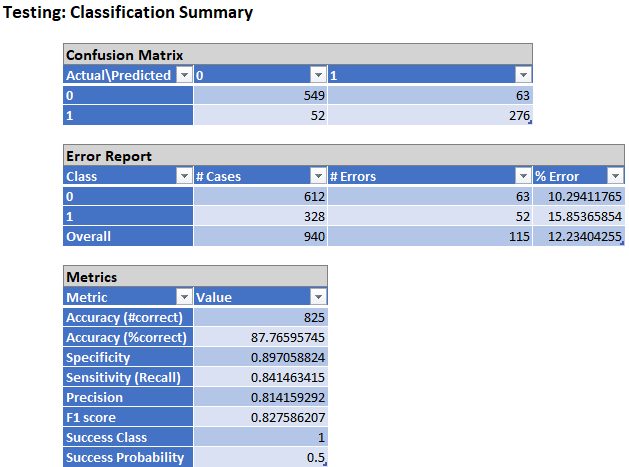
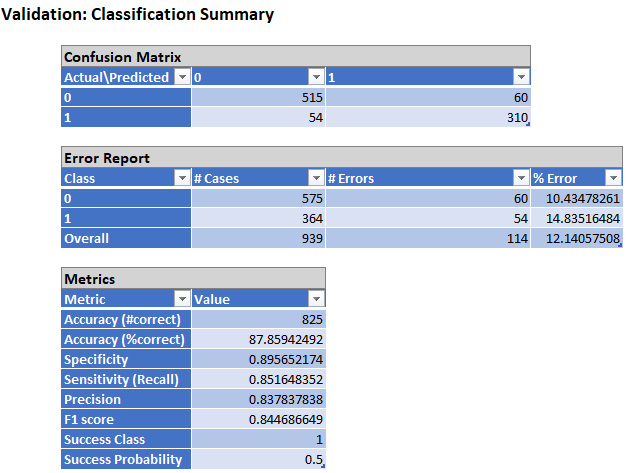
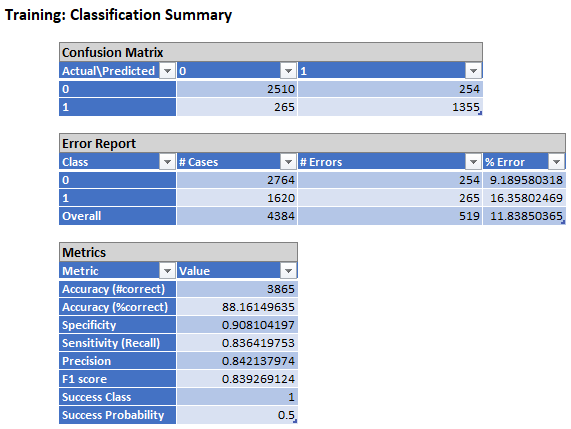
*Figure: Lift chart and ROC curve for model with AUROC = 0.9119*

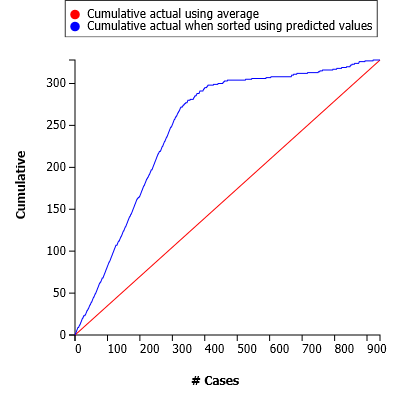
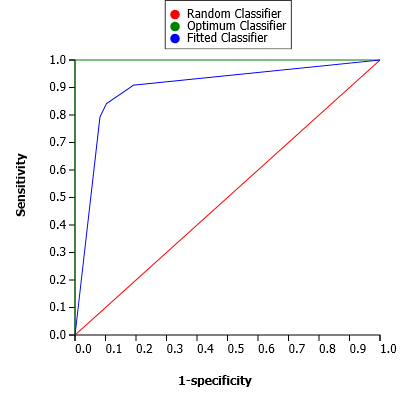
# Model 5 (response variable - spend\_all\_inchigh )

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

The model shows 11% error rate in training and 12% in validation and testing data sets. The model is performing average on testing sets with an accuracy of 87%.

*Table 23: Confusion matrix, error report and metrics for Model 5*





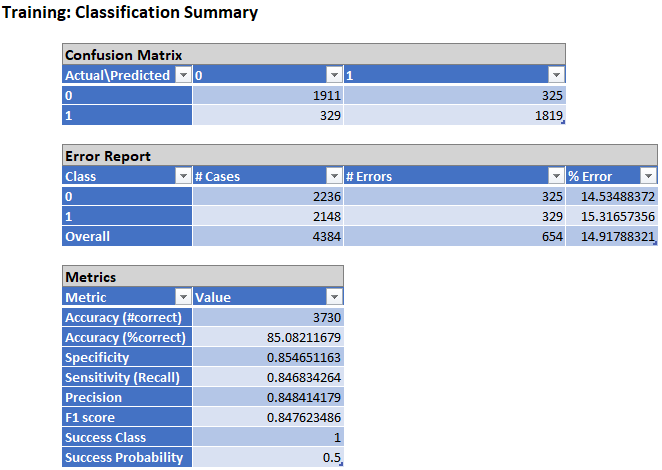
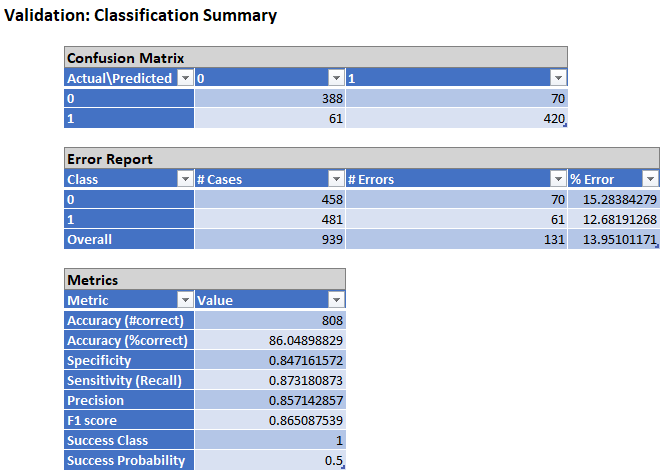
*Figure: Lift chart and ROC curve for model with AUROC = 0.8987*

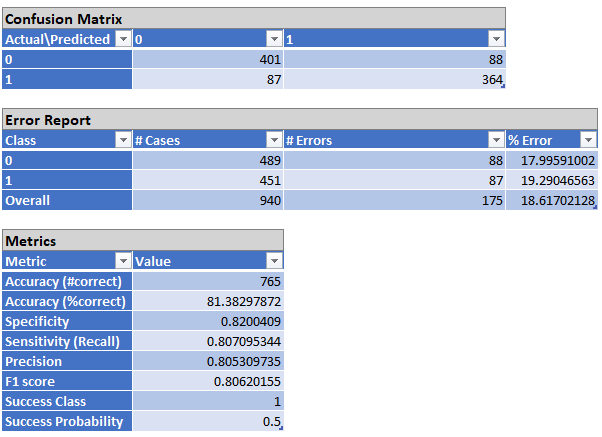
The AUROC for the model is 0.89, which shows the percentage of correctly classified records. Lift chart is not much more lifted than the red line representing the cumulative actual ––using average.

# Model 6 (response variable - spend\_all\_inmiddle )

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

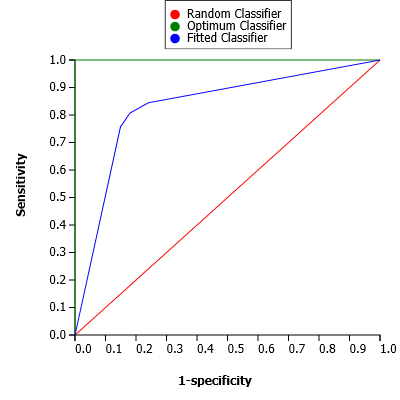
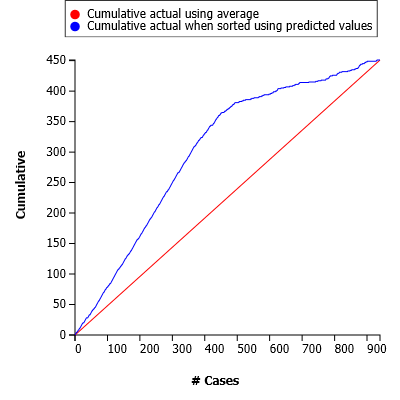
*Table 23: Confusion matrix, error report and metrics for Model 6*



The model shows 15% error rate in training, 14% in validation and 18% in testing data sets. This may be due to the overfitting of data. The model is performing average on testing sets with an accuracy of 81%.

The AUROC for the model is 0.8311, which shows the percentage of correctly classified records. Lift chart is not much more lifted than the red line representing the cumulative actual using average.

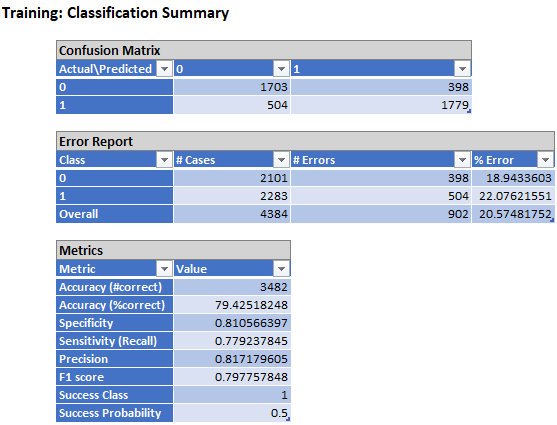
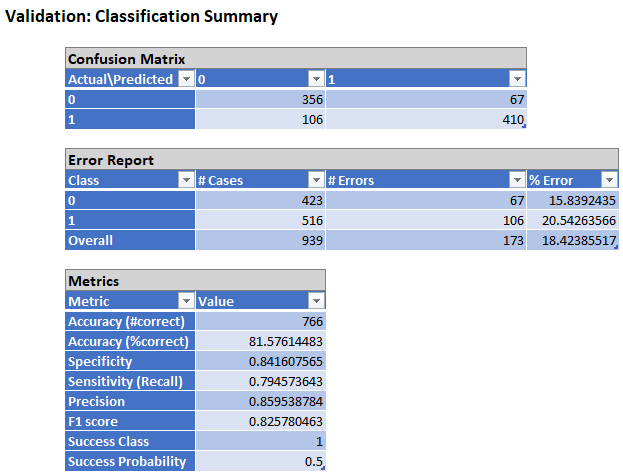


*Figure: Lift chart and ROC curve for model with AUROC = 0.8310*

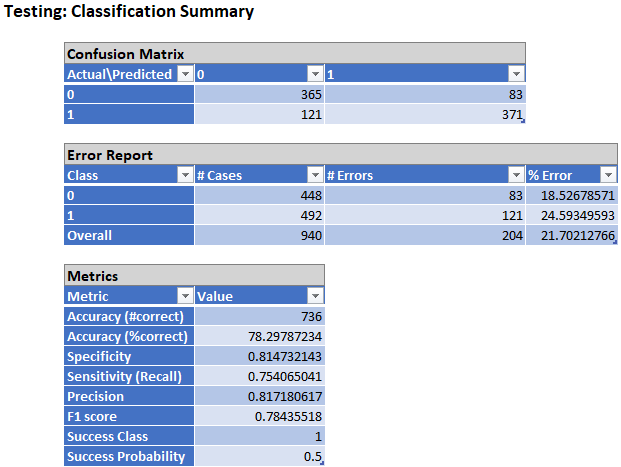
# Model 7 (response variable - spend\_all\_inclow )

The following table shows the confusion matrices, error reports, and performance metrics (accuracy, specificity, sensitivity, precision, and success class) for the model built.

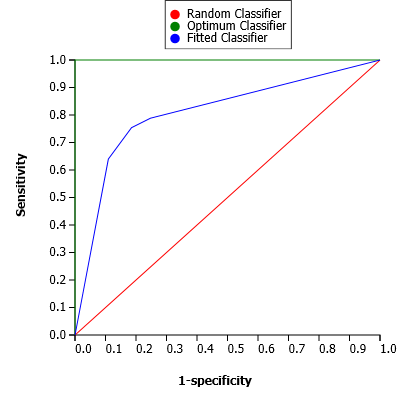
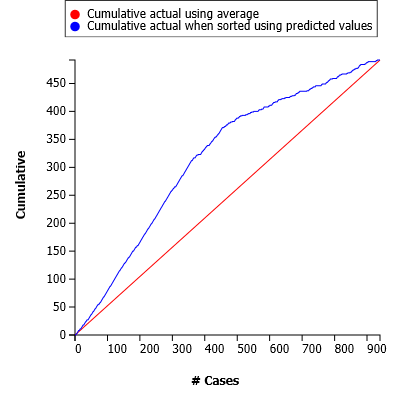
*Table 24: Confusion matrix, error report and metrics for Model 7*

The model shows 18% error rate for training and validation data sets.But, error rate increased to 22% for testing data sets. This may be due to the overfitting of data. The model is performing average on testing sets with an accuracy of 78%.



*Table 24: Confusion matrix, error report and metrics for Model 7*



*Figure: Lift chart and ROC curve for model with AUROC = 0.8088*

The AUROC for the model is 0.80, which shows the percentage of correctly classified records. Lift chart is not much more lifted than the red line representing the cumulative actual using average.

**Project Results:**

Based on the performance metrics of each model, the best model is Model 1(response variable spend\_acf). With the highest Area Under the ROC curve of 0.97, it proves to be the best classifier among the seven models. Additionally, it demonstrates a low error rate of 2%.

On the contrary, model 7 does not perform very well. It has high error of 21% rate among all seven models.

**Impact of the Project Outcomes:**

From the above models, we can observe that gps\_workplaces, gps\_away\_from\_home, and gps\_retail\_and\_recreation are important variables and have a high impact on consumer spending behavior.

The government can take action to improve safety at the places which affect consumer spending.

For example, if we consider Model-1 (response variable – spend\_acf) best-pruned tree, we can see that there is huge drop in consumer spending “If gps\_workplaces < -0.04 and if gps\_away\_from\_home < -0.05”.

Improper safety precautions (sanitization etc.) at workplaces might be one of the main reasons for this. So, by implementing the required safety precautions at the workplace, we can improve consumer spending on accommodation and food services.

*Figure: Model 1 best pruned tree*

