Mobility-Spending Classification

Project Report

Parikshith T (91)-9901665577

thriyambaka.p@northeastern.edu

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.

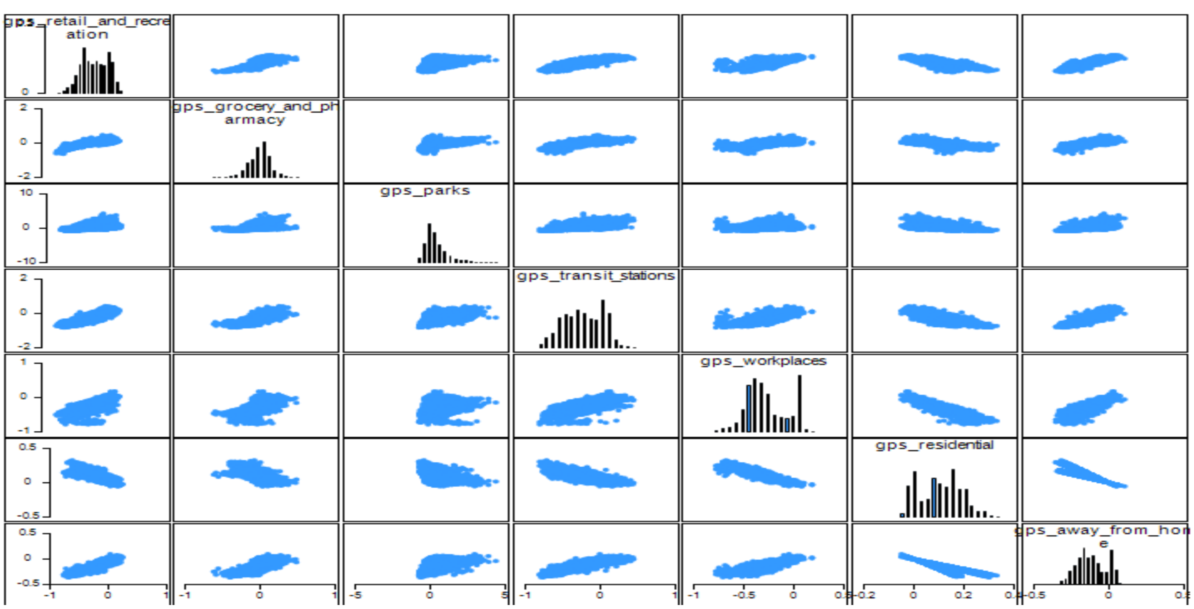

Figure 1: Scatterplot Matrix evaluating attribute correlation

Correlation matrix:

	gps_retail_and_recreation	gps_grocery_and_pharmacy	gps_parks	gps_transit_stations	gps_workplaces	gps_residential	gps_away_from_home
gps_retail_and_recreation	1						
gps_grocery_and_pharmacy	0.789776455	1					
gps_parks	0.352394315	0.460670368	1				
gps_transit_stations	0.886646162	0.778125473	0.370731	1			
gps_workplaces	0.788049293	0.613528891	0.1265962	0.803761754	1		
gps_residential	-0.831698431	-0.658886397	-0.2780845	-0.831174304	-0.94783379	1	
gps_away_from_home	0.929464671	0.744997489	0.3028153	0.88728299	0.900540016	-0.921951584	1

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.

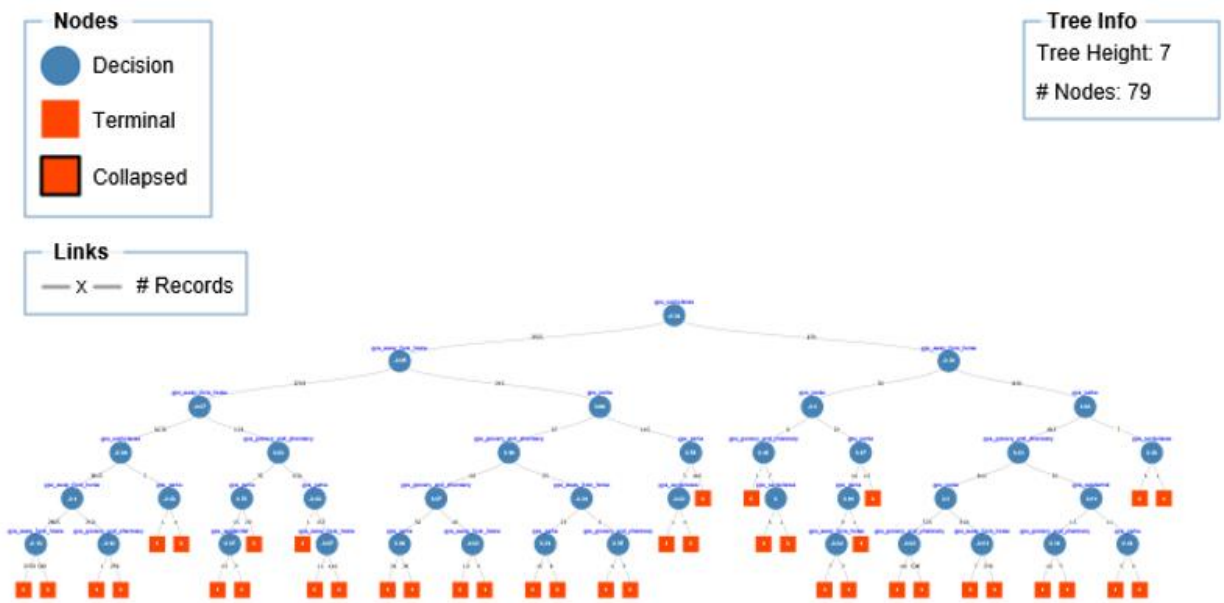

Figure 2: Fully grown tree

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.

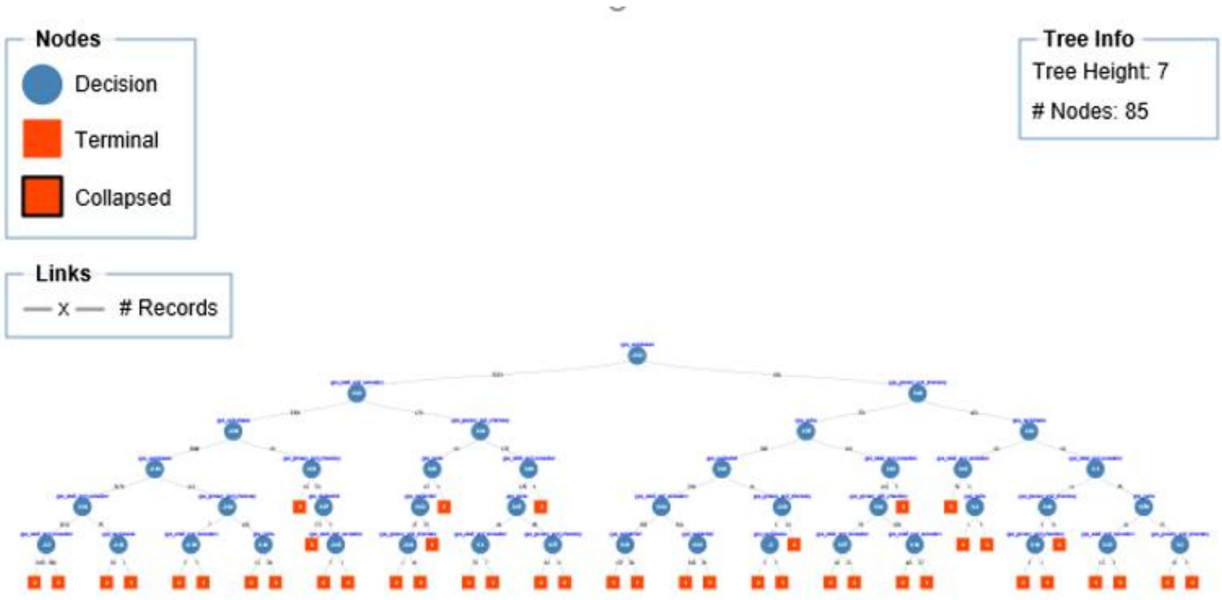

Figure: Fully grown tree

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.

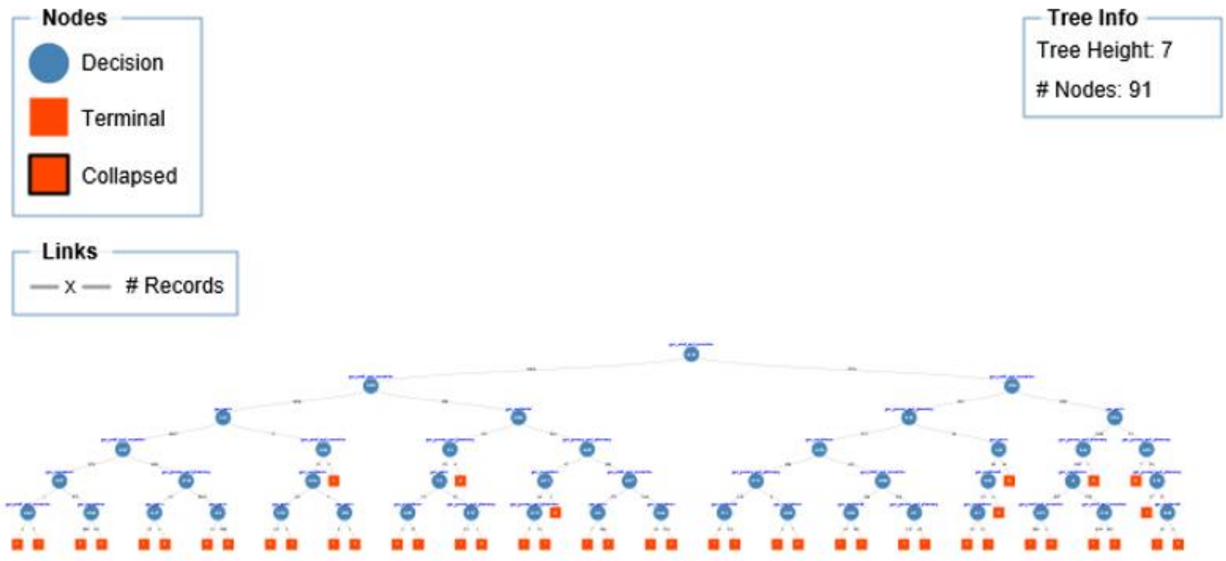

Figure: Fully grown tree

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.

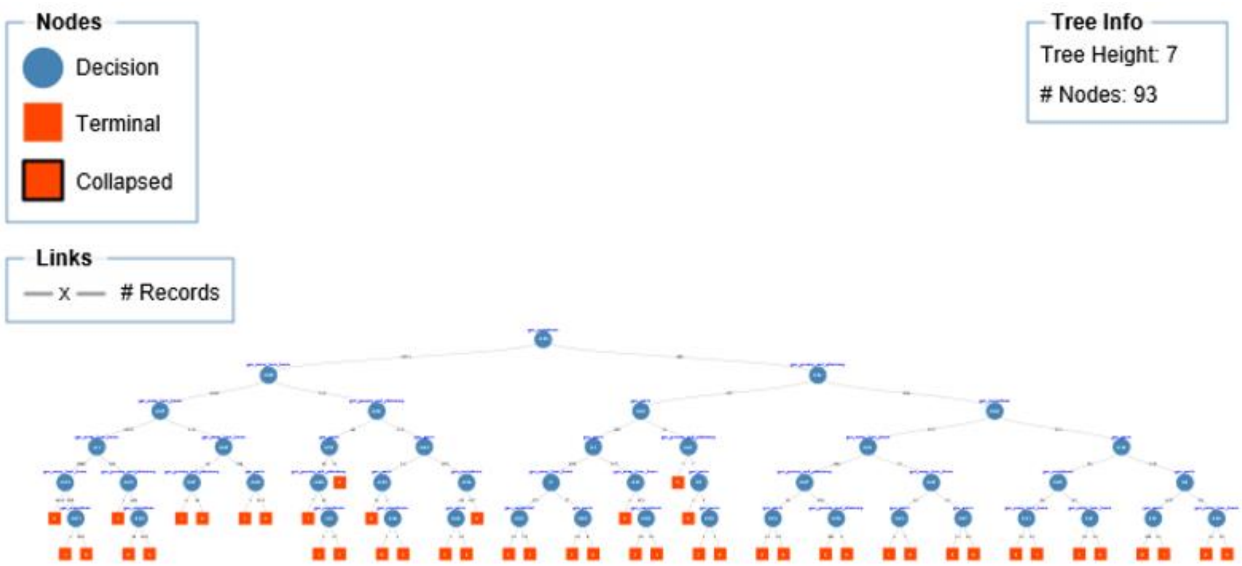

Figure: Fully grown tree

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.

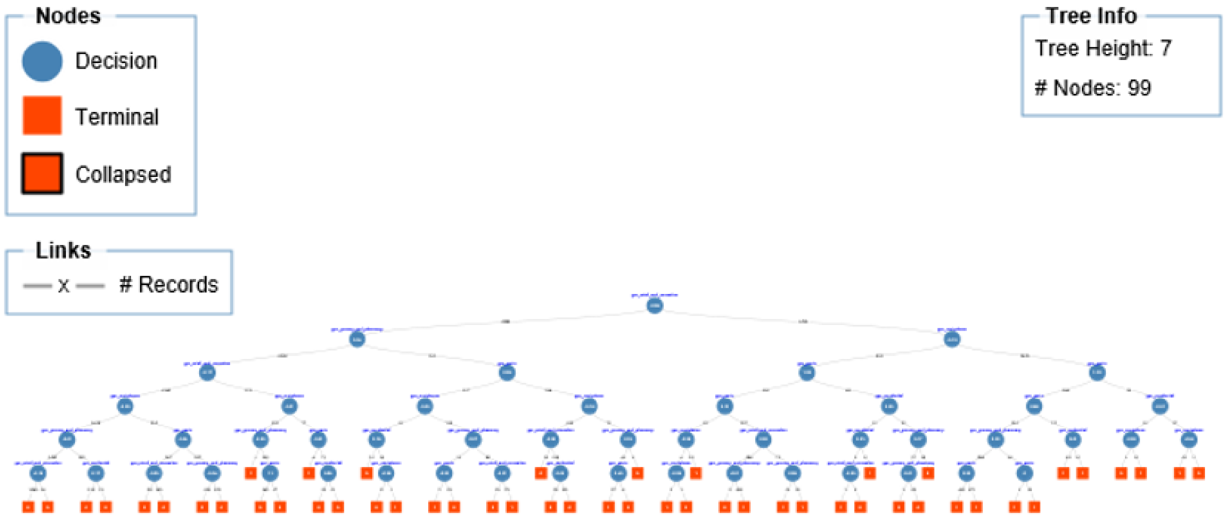

Figure: Fully grown tree

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.

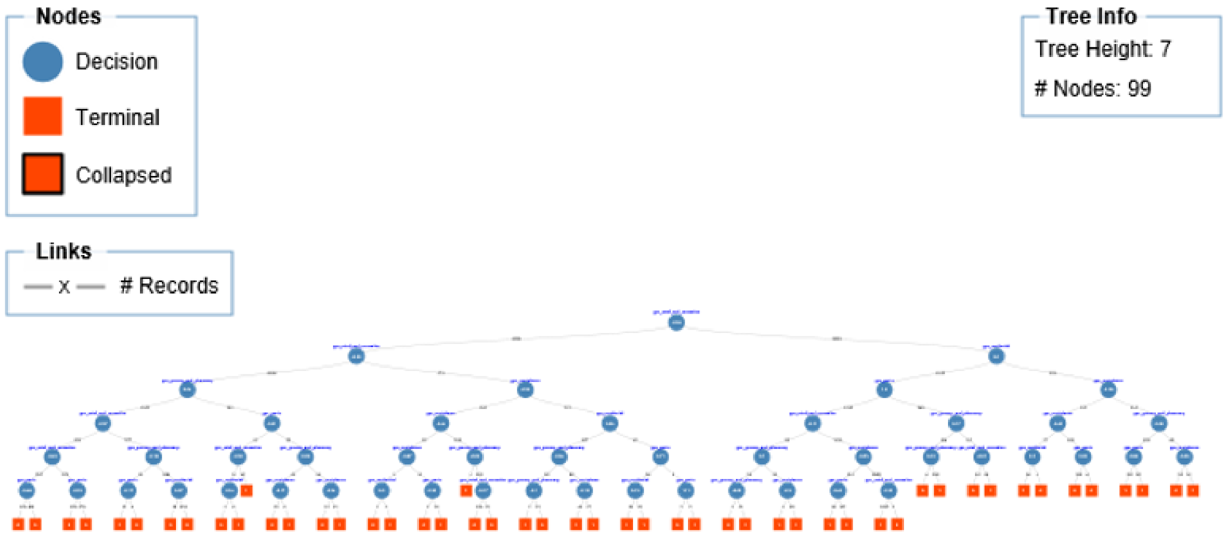

Figure: Fully grown tree

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.

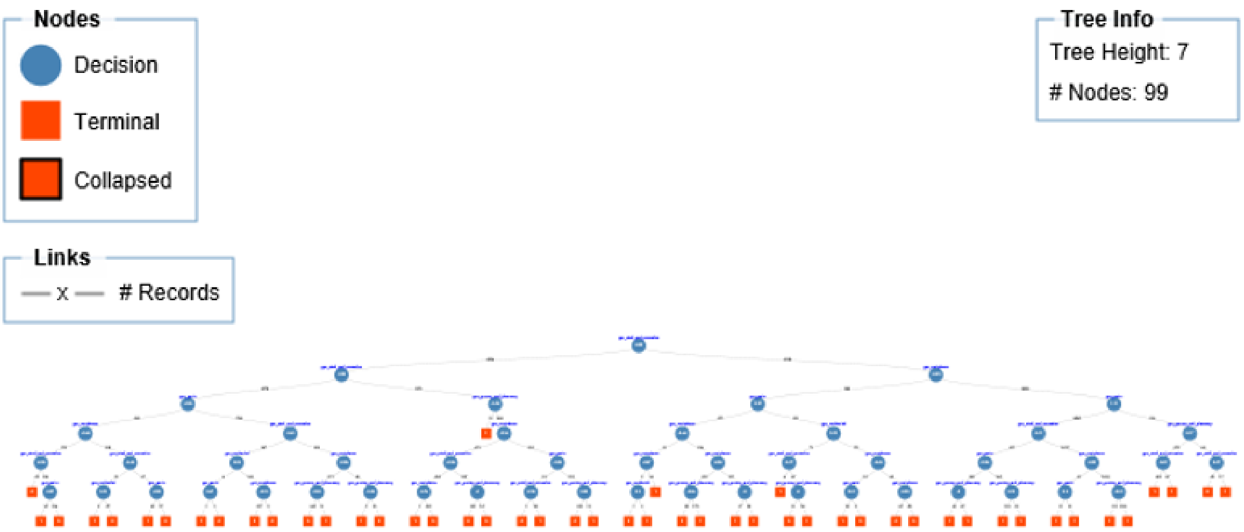

Figure: Fully grown tree

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

Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	3406	88
1	32	858

Error Report

Class	# Cases	# Errors	% Error
0	3494	88	2.51860332
1	890	32	3.595505618
Overall	4384	120	2.737226277

Metrics

Metric	Value
Accuracy (#correct)	4264
Accuracy (%correct)	97.2627737
Specificity	0.97481397
Sensitivity (Recall)	0.96404494
Precision	0.90697674
F1 score	0.93464052
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	727	17
1	9	186

Error Report

Class	# Cases	# Errors	% Error
0	744	17	2.284946237
1	195	9	4.615384615
Overall	939	26	2.768903088

Metrics

Metric	Value
Accuracy (#correct)	913
Accuracy (%correct)	97.23109691
Specificity	0.977150538
Sensitivity (Recall)	0.953846154
Precision	0.916256158
F1 score	0.934673367
Success Class	1
Success Probability	0.5

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	747	16
1	7	170

Error Report			
Class	# Cases	# Errors	% Error
0	763	16	2.096985583
1	177	7	3.95480226
Overall	940	23	2.446808511

Metrics	
Metric	Value
Accuracy (#correct)	917
Accuracy (%correct)	97.5531915
Specificity	0.97903014
Sensitivity (Recall)	0.96045198
Precision	0.91397849
F1 score	0.93663912
Success Class	1
Success Probability	0.5

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%.

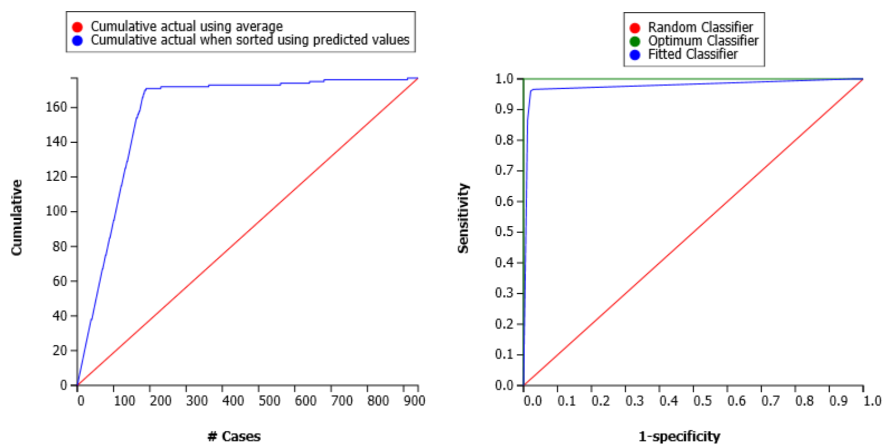

Figure: Lift chart and ROC curve for model with AUROC = 0.976

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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	3618	211
1	70	485

Error Report			
Class	# Cases	# Errors	% Error
0	3829	211	5.510577174
1	555	70	12.61261261
Overall	4384	281	6.409671533

Metrics	
Metric	Value
Accuracy (#correct)	4103
Accuracy (%correct)	93.5903285
Specificity	0.94489423
Sensitivity (Recall)	0.87387387
Precision	0.69683908
F1 score	0.7753797
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	769	56
1	25	89

Error Report			
Class	# Cases	# Errors	% Error
0	825	56	6.787878788
1	114	25	21.92982456
Overall	939	81	8.626198083

Metrics	
Metric	Value
Accuracy (#correct)	858
Accuracy (%correct)	91.37380192
Specificity	0.932121212
Sensitivity (Recall)	0.780701754
Precision	0.613793103
F1 score	0.687258687
Success Class	1
Success Probability	0.5

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	776	46
1	19	99

Error Report			
Class	# Cases	# Errors	% Error
0	822	46	5.596107056
1	118	19	16.10169492
Overall	940	65	6.914893617

Metrics	
Metric	Value
Accuracy (#correct)	875
Accuracy (%correct)	93.08510638
Specificity	0.944038929
Sensitivity (Recall)	0.838983051
Precision	0.682758621
F1 score	0.752851711
Success Class	1
Success Probability	0.5

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%.

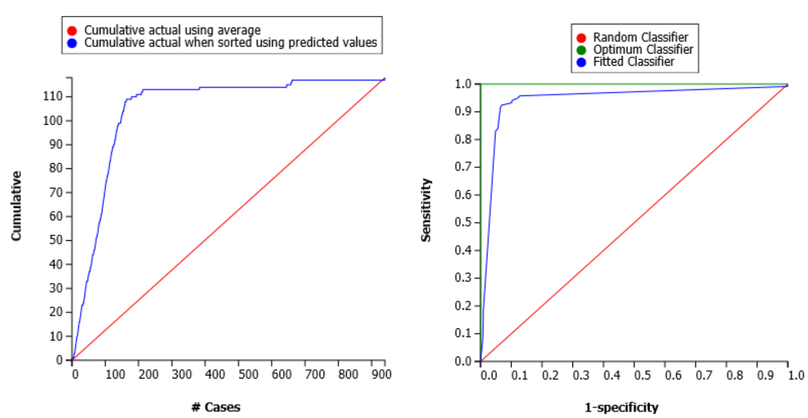

Figure: Lift chart and ROC curve for model with AUROC = 0.9441

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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	2357	207
1	254	1566

Error Report			
Class	# Cases	# Errors	% Error
0	2564	207	8.073322933
1	1820	254	13.95604396
Overall	4384	461	10.51551095

Metrics	
Metric	Value
Accuracy (#correct)	3923
Accuracy (%correct)	89.4844891
Specificity	0.91926677
Sensitivity (Recall)	0.86043956
Precision	0.88324873
F1 score	0.87169496
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	483	48
1	55	353

Error Report			
Class	# Cases	# Errors	% Error
0	531	48	9.039548023
1	408	55	13.48039216
Overall	939	103	10.96911608

Metrics	
Metric	Value
Accuracy (#correct)	836
Accuracy (%correct)	89.03088392
Specificity	0.90960452
Sensitivity (Recall)	0.865196078
Precision	0.880299252
F1 score	0.872682324
Success Class	1
Success Probability	0.5

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	511	50
1	65	314

Error Report			
Class	# Cases	# Errors	% Error
0	561	50	8.912655971
1	379	65	17.15039578
Overall	940	115	12.23404255

Metrics	
Metric	Value
Accuracy (#correct)	825
Accuracy (%correct)	87.7659574
Specificity	0.91087344
Sensitivity (Recall)	0.82849604
Precision	0.86263736
F1 score	0.84522207
Success Class	1
Success Probability	0.5

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%.

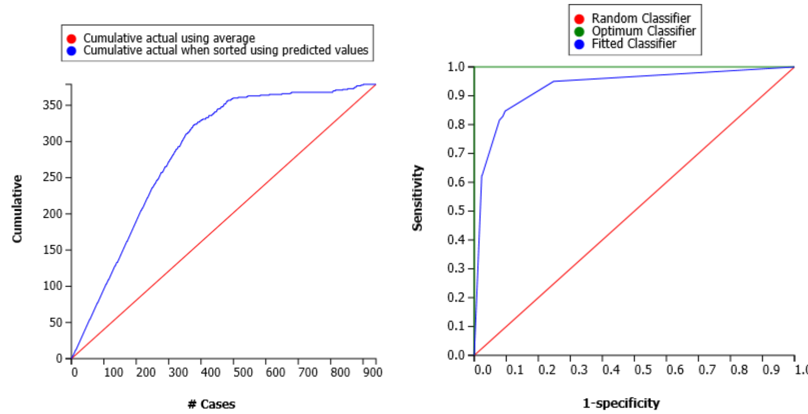

Figure: Lift chart and ROC curve for model with AUROC = 0.9311

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

Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	3839	125
1	188	232

Error Report

Class	# Cases	# Errors	% Error
0	3964	125	3.153380424
1	420	188	44.76190476
Overall	4384	313	7.13959854

Metrics

Metric	Value
Accuracy (#correct)	4071
Accuracy (%correct)	92.8604015
Specificity	0.9684662
Sensitivity (Recall)	0.55238095
Precision	0.64985994
F1 score	0.5971686
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	823	22
1	47	47

Error Report

Class	# Cases	# Errors	% Error
0	845	22	2.603550296
1	94	47	50
Overall	939	69	7.348242812

Metrics

Metric	Value
Accuracy (#correct)	870
Accuracy (%correct)	92.65175719
Specificity	0.973964497
Sensitivity (Recall)	0.5
Precision	0.68115942
F1 score	0.576687117
Success Class	1
Success Probability	0.5

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	836	22
1	34	48

Error Report			
Class	# Cases	# Errors	% Error
0	858	22	2.564102564
1	82	34	41.46341463
Overall	940	56	5.957446809

Metrics	
Metric	Value
Accuracy (#correct)	884
Accuracy (%correct)	94.04255319
Specificity	0.974358974
Sensitivity (Recall)	0.585365854
Precision	0.685714286
F1 score	0.631578947
Success Class	1
Success Probability	0.5

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.

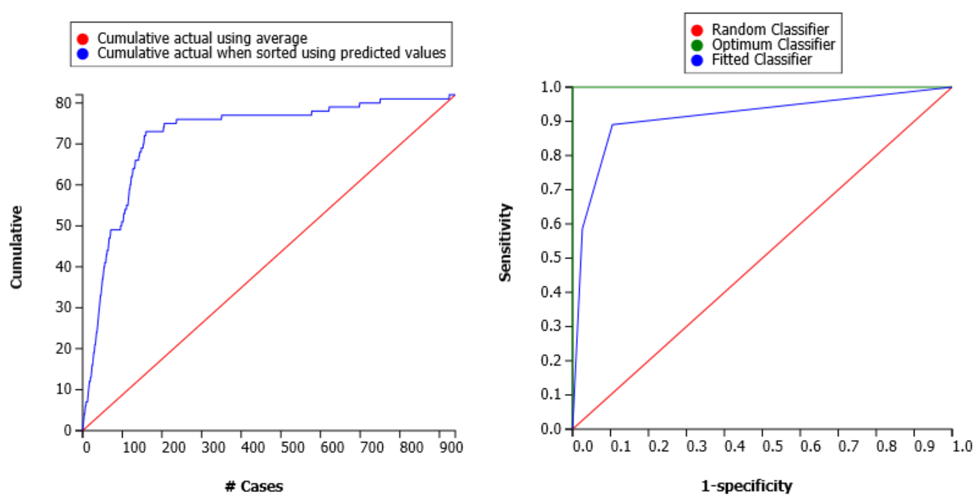

Figure: Lift chart and ROC curve for model with AUROC = 0.9119

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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	2510	254
1	265	1355

Error Report			
Class	# Cases	# Errors	% Error
0	2764	254	9.189580318
1	1620	265	16.35802469
Overall	4384	519	11.83850365

Metrics	
Metric	Value
Accuracy (#correct)	3865
Accuracy (%correct)	88.16149635
Specificity	0.908104197
Sensitivity (Recall)	0.836419753
Precision	0.842137974
F1 score	0.839269124
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	515	60
1	54	310

Error Report			
Class	# Cases	# Errors	% Error
0	575	60	10.43478261
1	364	54	14.83516484
Overall	939	114	12.14057508

Metrics	
Metric	Value
Accuracy (#correct)	825
Accuracy (%correct)	87.85942492
Specificity	0.895652174
Sensitivity (Recall)	0.851648352
Precision	0.837837838
F1 score	0.844686649
Success Class	1
Success Probability	0.5

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	549	63
1	52	276

Error Report			
Class	# Cases	# Errors	% Error
0	612	63	10.29411765
1	328	52	15.85365854
Overall	940	115	12.23404255

Metrics	
Metric	Value
Accuracy (#correct)	825
Accuracy (%correct)	87.76595745
Specificity	0.897058824
Sensitivity (Recall)	0.841463415
Precision	0.814159292
F1 score	0.827586207
Success Class	1
Success Probability	0.5

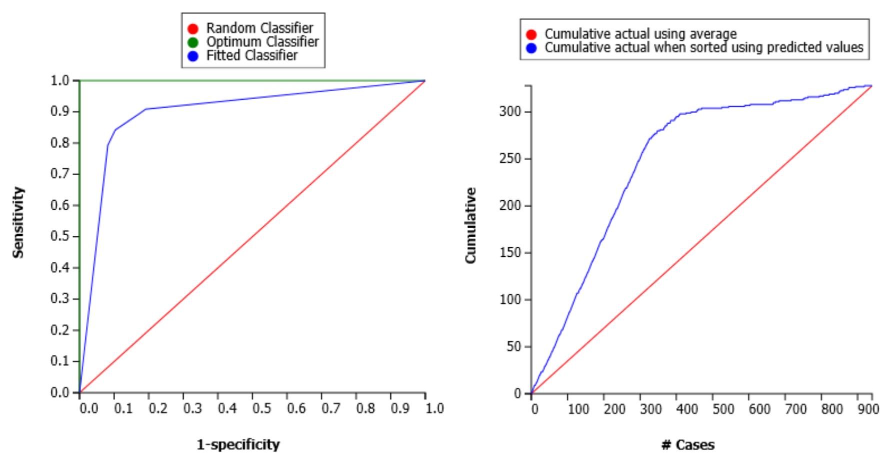

Figure: Lift chart and ROC curve for model with AUROC = 0.8987

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

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	1911	325
1	329	1819

Error Report			
Class	# Cases	# Errors	% Error
0	2236	325	14.53488372
1	2148	329	15.31657356
Overall	4384	654	14.91788321

Metrics	
Metric	Value
Accuracy (#correct)	3730
Accuracy (%correct)	85.08211679
Specificity	0.854651163
Sensitivity (Recall)	0.846834264
Precision	0.848414179
F1 score	0.847623486
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	388	70
1	61	420

Error Report			
Class	# Cases	# Errors	% Error
0	458	70	15.28384279
1	481	61	12.68191268
Overall	939	131	13.95101171

Metrics	
Metric	Value
Accuracy (#correct)	808
Accuracy (%correct)	86.04898829
Specificity	0.847161572
Sensitivity (Recall)	0.873180873
Precision	0.857142857
F1 score	0.865087539
Success Class	1
Success Probability	0.5

Confusion Matrix		
Actual\Predicted	0	1
0	401	88
1	87	364

Error Report			
Class	# Cases	# Errors	% Error
0	489	88	17.99591002
1	451	87	19.29046563
Overall	940	175	18.61702128

Metrics	
Metric	Value
Accuracy (#correct)	765
Accuracy (%correct)	81.38297872
Specificity	0.8200409
Sensitivity (Recall)	0.807095344
Precision	0.805309735
F1 score	0.80620155
Success Class	1
Success Probability	0.5

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.

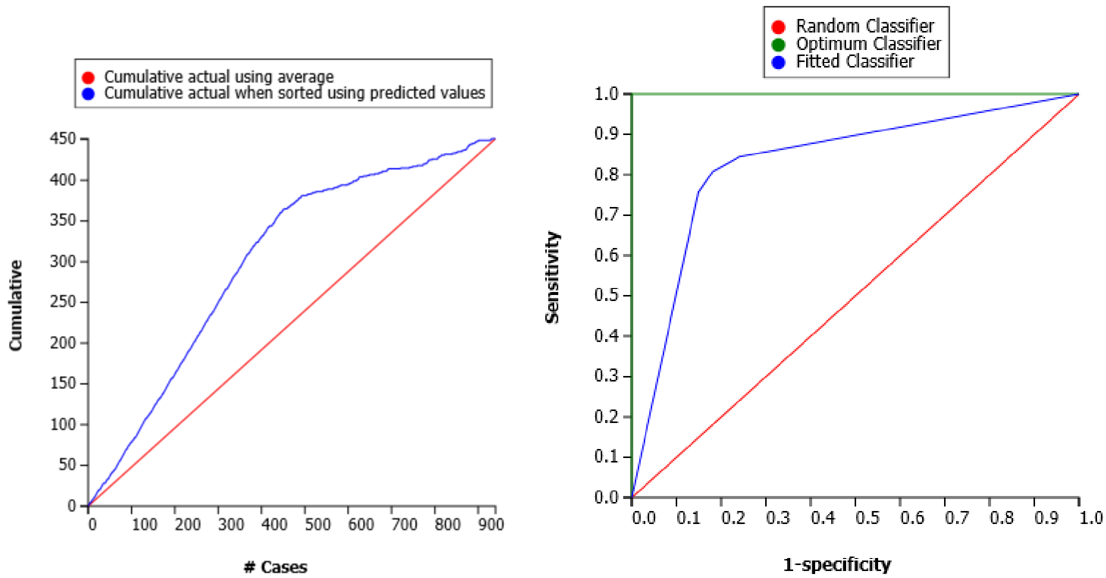

Figure: Lift chart and ROC curve for model with AUROC = 0.8310

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

Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	1703	398
1	504	1779

Error Report

Class	# Cases	# Errors	% Error
0	2101	398	18.9433603
1	2283	504	22.07621551
Overall	4384	902	20.57481752

Metrics

Metric	Value
Accuracy (#correct)	3482
Accuracy (%correct)	79.42518248
Specificity	0.810566397
Sensitivity (Recall)	0.779237845
Precision	0.817179605
F1 score	0.797757848
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	356	67
1	106	410

Error Report

Class	# Cases	# Errors	% Error
0	423	67	15.8392435
1	516	106	20.54263566
Overall	939	173	18.42385517

Metrics

Metric	Value
Accuracy (#correct)	766
Accuracy (%correct)	81.57614483
Specificity	0.841607565
Sensitivity (Recall)	0.794573643
Precision	0.859538784
F1 score	0.825780463
Success Class	1
Success Probability	0.5

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%.

Testing: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	365	83
1	121	371

Error Report			
Class	# Cases	# Errors	% Error
0	448	83	18.52678571
1	492	121	24.59349593
Overall	940	204	21.70212766

Metrics	
Metric	Value
Accuracy (#correct)	736
Accuracy (%correct)	78.29787234
Specificity	0.814732143
Sensitivity (Recall)	0.754065041
Precision	0.817180617
F1 score	0.78435518
Success Class	1
Success Probability	0.5

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

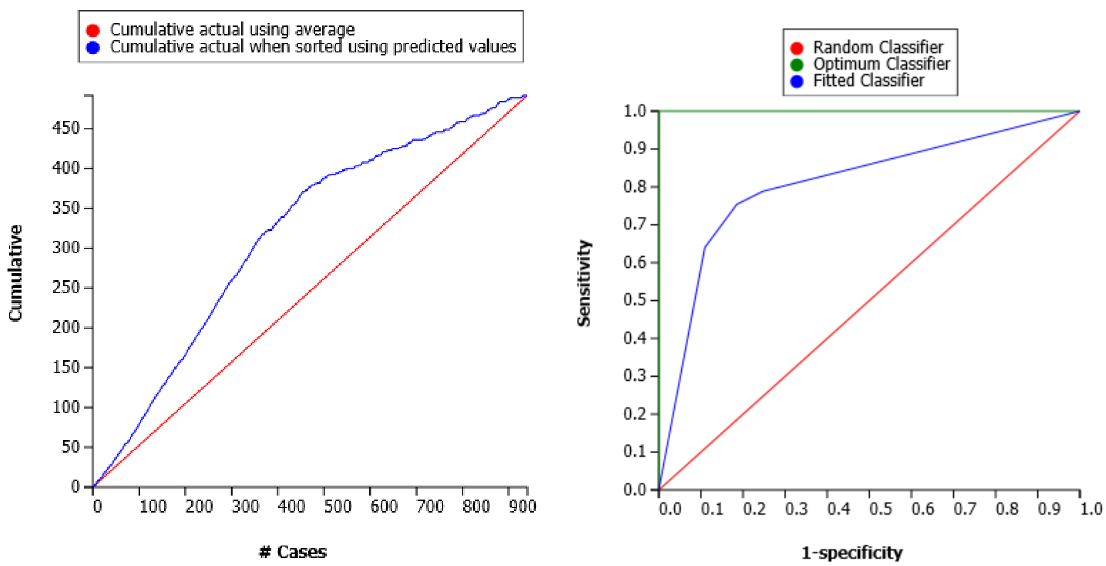

Figure: Lift chart and ROC curve for model with AUROC = 0.8088

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