Hyundai Retention Service Data

Archit Jajoo

Marielle Garcia

Edward Chang

Flip Pua

Debbie Wang

Ken Poon

Jeff Reyes

California State University, Fullerton

ISDS 574 – Data Mining for Managers

Dr. Yinfei Kong

May 14, 2017

Executive Summary

Less than half of first year owners return to a Hyundai dealership for routine car maintenance at least twice per year and less than a third remain repeat service customers by third ownership year. Repeat service customers are more likely to buy another car from Hyundai in the future, so Hyundai aims to increase sales of new cars through service retention.

This project is focused on data mining the customer database of Hyundai Motor America by developing a classification model for service retention. The primary goal is to predict which recent service customers are likely to become repeat customers and understand the characteristics of these individuals.

Hyundai has millions of service customers in the United States since it sold the first car in 1986. But, only service customers from last year that completed all customer surveys are included in the dataset due to XLMiner's limit of 10,000 records. After sampling, data cleansing and dimension reduction, the final dataset consists of 4,072 records with 50% repeat customers, 50% not repeat customers, 15 continuous variables, 8 ordinal variables and 18 binary dummy variables.

kNN, Logistic Regression and CART models developed and evaluated. The best pruned tree had 14 decision nodes with Customer Paid RO Dollars as the first variable split. Logistic regression model used backward, forward and stepwise variable selection methods. And, kNN model yielded a best k value of 17. Ultimately, logistic regression model with five selected variables based on forward/stepwise selection yielded the best performing classification model with a misclassification rate of 22.96% and Logit = 2.5937 + 0.0002MileageAtLastService + 0.8214NumberofInCabinAirFiltersPurchased – 0.2248VehicleAgeInMonths + 0.1143AverageMonthsBetweenPaidServiceVisits –

0.0004AverageMilesBetweenPaidServiceVisit. This means Hyundai owners are more likely to be repeat customers when they purchase more cabin air filters and have longer service intervals but tend to go elsewhere for maintenance the older the car gets.

Using this information, Hyundai will target each customer type with different push/pull marketing tactics, such as cabin air filter specials and time specific service reminders with more attractive discounts and offers. It may be possible to expand this analysis to include older customers and other service time periods by using more capable and user friendly analytic software like SAS or Alteryx.

Section 1: Introduction: background and problem description Hyundai Motor America has been selling cars in the United States since 1986. By the end of 2015, Hyundai Motor America posted its best year ever with sales of 761,710 Hyundai vehicles. While Hyundai sales performed well, the company still struggles with getting most of its owners to come back to a Hyundai dealership for routine car maintenance. 48% of the cars retailed in 2015 returned to a Hyundai dealership for routine car maintenance at least twice in 2016 and 32% of the people who bought a car in 2013 returned to a Hyundai dealership for routine car maintenance at least twice in 2016. Studies show Hyundai owners who continue to service their car at least twice per year at a Hyundai dealership are 33% more likely buy another Hyundai car in the next 10 years, so Hyundai aims to increase sales of new cars through service retention. With better practices in place, Hyundai can ideally increase its customer retention rate from 48% to 80% in Year 1, 40% to 70% in Year 2 and 32% to 60% in Year 3, which will yield more loyal customers repurchasing another new Hyundai car in the long run.

For this project, a customer who returns at least twice per year for routine maintenance is defined as a “repeat service customer.” This project is focused on data mining the customer database of Hyundai Motor America by running a sample dataset through classification models using CART, kNN and logistic regression methods, selecting the best model for identifying repeat service customers and understanding the best model’s results. The primary goal is to predict which recent service customers are likely to become repeat customers or not repeat customers and understand the characteristics of these customers to better serve their needs and earn their business.

Section 2: Questions of interest, description of the variables in the study

Using Hyundai Service Retention data, we want to answer the question of whether a Hyundai car owner is likely to return to a Hyundai dealership for car maintenance at least twice per year and become a "repeat Hyundai service customer." This is a classification question with 2 outcomes: 1-repeat Hyundai service customer or 0-not a repeat Hyundai service customer. Identifying a customer as potential repeat customer or potential defector at an early stage of customer lifecycle will enable Hyundai to address each customer type differently by fostering repeat business from valuable customers and working harder to recover potential defectors before they develop a strong affinity to do business with a competing car maintenance provider.

Another purpose for this analysis is to interpret the data model to find out what are significant predictors of each outcome and understand which factors best help classify a repeat Hyundai service customer and which factors do not help in this classification. Understanding which key performance indicators, programs and initiatives have a higher impact on customer retention will enable Hyundai to better focus its efforts and investments to encourage more repeat business from its customers.

Section 3: Data pre-processing and exploration

Hyundai Motor America's database includes millions of customer transactions and records. After pre-processing and exploration, the final cleaned up dataset included 4,072 records with 15 continuous variables, 8 ordinal variables and 18 binary dummy variables, and 1 outcome variable 'Repeat Customer' (50% repeat customer and 50% not repeat customer). The sample data contains last year's (2016) service customers in their first and second year of new vehicle ownership and is sourced from Hyundai Motor America’s operational data stores. Excel with XLMiner were used to prepare and model the data.

The following pre-processing steps were performed:

1. **Need for sampling – random or “selective”?**

   This project used a combination of mostly selective with random sampling method. Hyundai Motor America currently generates over five million dealership maintenance service transactions in a year, but XLMiner’s data mining capacity is limited to 10,000 records. Domain knowledge determines that it is more important to take care of new custome\rs and increase retention rate over the long run to maximize customer lifetime value; consequently, samples were selected from an initial group of over 600,000 first- and second-year customer records from customer-paid service transactions in the year 2016.

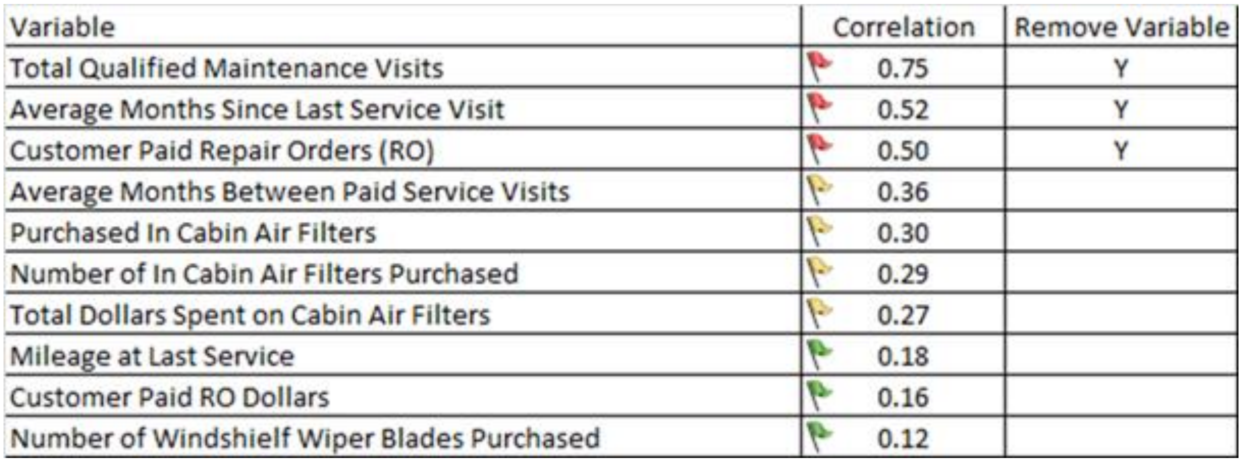
   These customer records were further filtered to include only customers that responded to all three types of Hyundai customer feedback surveys (Sales Satisfaction, Initial Quality Satisfaction, and Service Satisfaction) in order to reduce the records to an amount XLMiner can handle. Applying these filters yielded about 10,000 records, but 80% of the records were repeat customers, which caused the CART and logistic regression models not to develop normally. From the approximately 10,000 filtered records, all 2,036 non-repeat customers and 2,036 randomly selected repeat customers were included in the final data set. This final sample of 4,072 records ultimately enabled the models to work with a data set that is 50% repeat and 50% non-repeat customers.

2. **Specification of the “relevant” variables (do you need all of them?)**

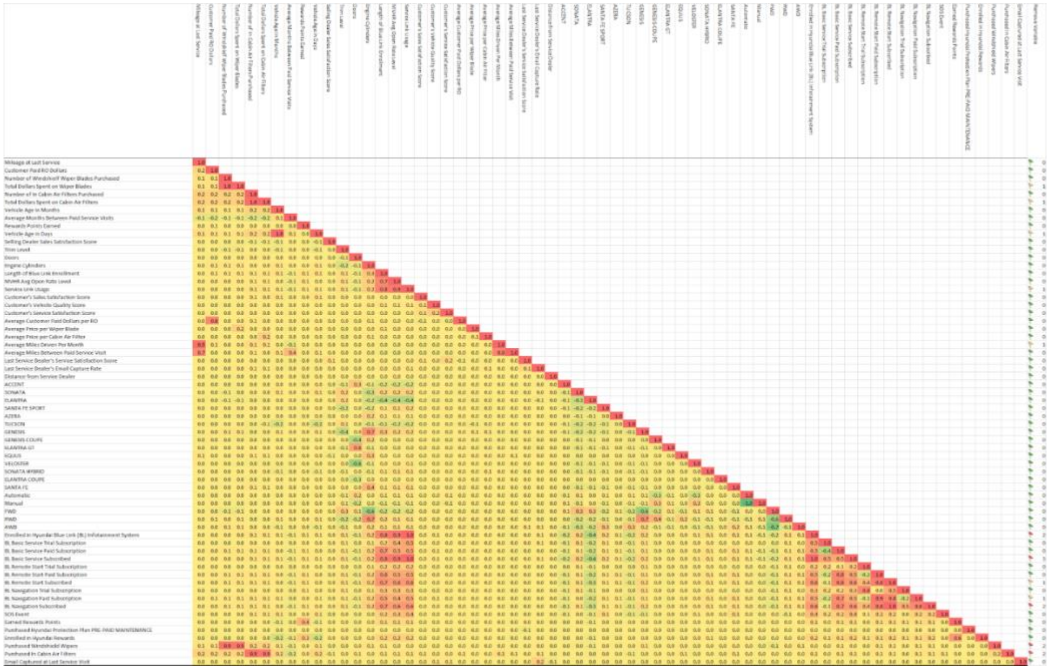
   Initially, domain knowledge and our questions prompted the specification of relevant variables to include in this data mining exercise. If Hyundai wants to know whether a program (e.g., Hyundai Rewards or Hyundai Protection Plan Pre-paid Maintenance) or a factor (e.g., customer satisfaction, price and distance) has an effect on service retention, then it is important that those related variables be included in the initial selected variables.

However, Principal Component Analysis indicates as few as 31 principal components explain over 95% of the variance. This implies multicollinearity or strong correlation among many independent variables and a need for data reduction. Correlation analysis showed three variables to be too strongly correlated to the outcome variable, so those three input variables were removed. Correlation analysis also showed some input variables have strong correlation with multiple other input variables, were also removed.

Please see below and appendix for correlation analysis and removed variables.



Variable	Correlation	Remove Variable
Total Qualified Maintenance Visits	0.75	Y
Average Months Since Last Service Visit	0.52	Y
Customer Paid Repair Orders (RO)	0.50	Y
Average Months Between Paid Service Visits	0.36	
Purchased In Cabin Air Filters	0.30	
Number of In Cabin Air Filters Purchased	0.29	
Total Dollars Spent on Cabin Air Filters	0.27	
Mileage at Last Service	0.18	
Customer Paid RO Dollars	0.16	
Number of Windshielf Wiper Blades Purchased	0.12	



3. Detecting missing values and outliers

The above mentioned data sampling procedure utilized missing value detection to delete records where any sales, product, or service satisfaction survey was missing. Domain knowledge and a scan of the minimum and maximum values from each variable enabled the use of outlier detection. Some transactional variables with missing values, such as customers who did not purchase a cabin air filter or prepaid maintenance package, were replaced with 0. While some other average price variables with missing values used imputed average prices of wipers and cabin air filters in order to not skew the price lower.

4. Use of visualization methods and summary statistics

We summarized results in Table 1 and include high-level or important plots as deemed necessary. See appendix.

Table 1 contains the description of the final cleaned dataset of 42 variables (1 outcome, 15 continuous, 8 ordinal and 18 binary) used by the data mining models:

Table 1

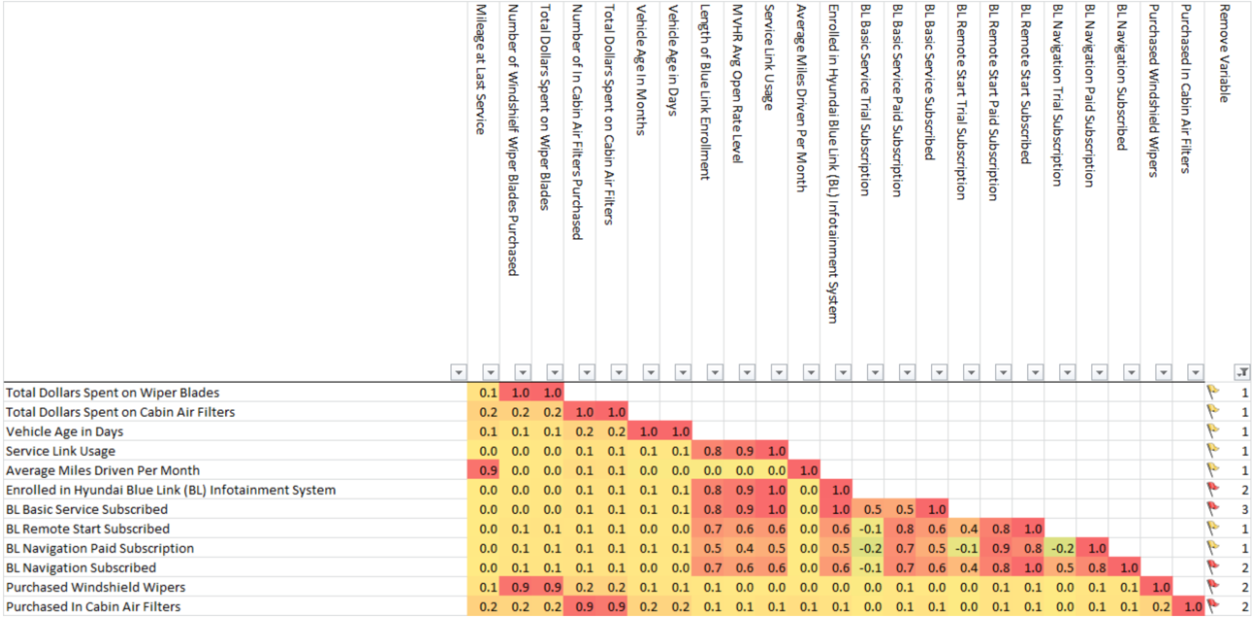
Variable Name	**Description**	**Type**	**Descriptive Statistics**
Repeat Customer	Outcome Variable 1- A Hyundai service customer who had two or more qualified service visits at a Hyundai dealership in the last 12 months of vehicle ownership 0- A Hyundai service customer who had only one qualified service visit at a Hyundai dealership in the last 12 months of vehicle ownership A qualified service visit is defined as a service where Hyundai oil filter is installed or customer paid $5 or more for a service visit	Binary	1: 2,036 (50%) 0: 2,036 (50%)
Mileage at Last Service	Odometer mileage reading of customer's vehicle during last service visit	Continuous	Avg: 15,650.09 StdDev: 16,439.14
Customer Paid RO Dollars	Total amount of dollars paid by customer on service repair orders (RO's)	Continuous	Avg: 248.03 StdDev: 607.2
Number of Windshield Wiper Blades Purchased	Number of windshield wipers purchased by customer from dealership	Continuous	Avg: 0.15 StdDev: 0.61
Number of In Cabin Air Filters Purchased	Number of in cabin air filters purchased by customer from dealership	Continuous	Avg: 0.29 StdDev: 0.54
Vehicle Age In Months	Customer's vehicle age in months	Continuous	Avg: 19.28 StdDev: 3.43
Average Months Between Paid Service Visits	Average months between customer paid service visits. If customer only had one visit where customer paid for service, then this is the average number of months from new car purchase to first paid service visit	Continuous	Avg: 5.74 StdDev: 3.54
Rewards Points Earned	Total points issued to customer through Hyundai Rewards loyalty program	Continuous	Avg: 234.72 StdDev: 1,401.23
Selling Dealer Sales Satisfaction Score	Sales satisfaction score (on a 100 to 1000 point scale) of the dealer where customer purchased their new Hyundai	Continuous	Avg: 953.84 StdDev: 53.59
Length of Blue Link Enrollment	Number of months customers enrolled in Hyundai Blue Link connected car (telematics) services	Continuous	Avg: 9.19 StdDev: 9.21
Average Customer Paid Dollars per RO	Average customer paid dollars per repair order (RO)	Continuous	Avg: 76.21

			StdDev: 286.16
Average Price per Wiper Blade	Average price per wiper blade sold at Hyundai dealership	Continuous	Avg: 16.05 StdDev: 1.07
Average Price per Cabin Air Filter	Average price per cabin air filter sold at Hyundai dealership	Continuous	Avg: 24.72 StdDev: 3.03
Average Miles Between Paid Service Visit	Average miles a customer drives between customer paid service visits	Continuous	Avg: 5,576.24 StdDev: 5,844.97
Last Service Dealer's Email Capture Rate	The percent of customer email addresses captured by the dealership where customer last serviced his or her vehicle	Continuous	Avg: 0.71 StdDev: 0.15
Distance from Service Dealer	The distance in miles a owner of vehicle resides from the nearest Hyundai dealership	Continuous	Avg: 21.55 StdDev: 102.72
Trim Level	Different grades or versions of the same model with different features and equipment. 1 represents base model and higher values represent more features or limited model	Ordinal	Avg: 3.19 StdDev: 2.55
Doors	Number of doors on the customer's vehicle	Ordinal	Avg: 3.99 StdDev: 0.32
Engine Cylinders	Number of engine cylinders in customer's vehicle	Ordinal	Avg: 4.38 StdDev: 0.85
MVHR Avg Open Rate Level	Monthly Vehicle Health Report (MVHR) email open rate level on 1 to 5 point scale	Ordinal	Avg: 1.39 StdDev: 1.27
Customer's Sales Satisfaction Score	Customer's satisfaction score of their new vehicle sales experience (normalized to 1 to 5 point scale)	Ordinal	Avg: 4.87 StdDev: 0.39
Customer's Vehicle Quality Score	Customer's satisfaction score of their new vehicle's initial product quality (normalized to 1 to 5 point scale)	Ordinal	Avg: 4.48 StdDev: 0.84
Customer's Service Satisfaction Score	Customer's satisfaction score of their overall dealership repair or maintenance service experience (normalized to 1 to 5 point scale)	Ordinal	Avg: 4.54 StdDev: 0.81
Last Service Dealer's Service Satisfaction Score	The overall service satisfaction score (on 1 to 5 point scale) of dealership where a customer last serviced his or her vehicle	Ordinal	Avg: 4.62 StdDev: 0.14
SANTA FE SPORT	Customer drives a Hyundai Santa Fe Sport model (1 = Yes, 0 = No)	Binary	1: 477 (12%) 0: 3,595 (88%)
TUCSON	Customer drives a Hyundai Tuscon model (1 = Yes, 0 = No)	Binary	1: 413 (10%)

			0: 3,659 (90%)
GENESIS	Customer drives a Hyundai Genesis model (1 = Yes, 0 = No)	Binary	1: 396 (10%) 0: 3,676 (90%)
OTHER MODELS	Customer drives any other Hyundai model that is not listed above or Sonata (1 = Yes, 0 = No)	Binary	1: 836 (21%) 0: 3,236 (79%)
Manual	Customer's vehicle is equipped with manual transmission type (1 = Yes, 0 = No)	Binary	1: 104 (3%) 0: 3,968 (97%)
RWD	Customer's vehicle is equipped with rear wheel drive (RWD) (1 = Yes, 0 = No)	Binary	1: 358 (9%) 0: 3,714 (91%)
AWD	Customer vehicle is equipped with all wheel drive (AWD) (1 = Yes, 0 = No)	Binary	1: 591 (15%) 0: 3,481 (85%)
BL Basic Service Trial Subscription	Customer subscribed to Blue Link (BL) basic connected car trial service (1 = Yes, 0 = No)	Binary	1: 1,296 (32%) 0: 2,776 (68%)
BL Basic Service Paid Subscription	Customer subscribed to Blue Link (BL) basic connected car paid service (1 = Yes, 0 = No)	Binary	1: 1,200 (29%) 0: 2,872 (71%)
BL Remote Start Trial Subscription	Customer subscribed to Blue Link (BL) connected car trial service with remote start package (1 = Yes, 0 = No)	Binary	1: 356 (9%) 0: 2,856 (91%)
BL Remote Start Paid Subscription	Customer subscribed to Blue Link (BL) connected car paid service with remote start package (1 = Yes, 0 = No)	Binary	1: 1,216 (30%) 0: 3,578 (70%)
BL Navigation Trial Subscription	Customer subscribed to Blue Link (BL) connected car paid service with navigation package (1 = Yes, 0 = No)	Binary	1: 494 (12%) 0: 3,376 (88%)
SOS Event	Customer initiated an SOS distress call from their car using Blue Link connected car service (1 = Yes, 0 = No)	Binary	1: 696 (17%) 0: 3,499 (83%)
Earned Rewards Points	Customer has earned points through Hyundai Rewards loyalty program (1 = Yes, 0 = No)	Binary	1: 573 (14%) 0: 3,499 (86%)
Purchased Hyundai Protection Plan PRE-PAID MAINTENANCE	Customer purchased a Hyundai Protection Plan branded prepaid maintenance package from dealership (1 = Yes, 0 = No)	Binary	1: 83 (2%) 0: 3,989 (98%)
Enrolled in Hyundai Rewards	Customer has enrolled in Hyundai Rewards loyalty program (1 = Yes, 0 = No)	Binary	1: 1,160 (28%) 0: 2,912 (72%)
Email Captured at Last Service Visit	Customer's email was recorded by dealership on customer's last repair order (1 = Yes, 0 = No)	Binary	1: 3,558 (87%) 0: 514 (13%)

Section 4: Dimension reduction/aggregation

Principal Component Analysis indicated 95% of variance is explained by potentially 30 variables, so correlation analysis is used to identify redundant variables and reduced the number of variables included in the final data set. Please see below and appendix for correlation matrices and which input variables were removed due to multicollinearity and high correlation to outcome variable.



	Mileage at Last Service	Number of Windshielf Wiper Blades Purchased	Total Dollars Spent on Wiper Blades	Number of In Cabin Air Filters Purchased	Total Dollars Spent on Cabin Air Filters	Vehicle Age in Months	Vehicle Age in Days	Length of Blue Link Enrollment	MVHR Avg Open Rate Level	Service Link Usage	Average Miles Driven Per Month	Enrolled in Hyundai Blue Link (BL) Infotainment System	BL Basic Service Trial Subscription	BL Basic Service Paid Subscription	BL Basic Service Subscribed	BL Remote Start Trial Subscription	BL Remote Start Paid Subscription	BL Remote Start Subscribed	BL Navigation Trial Subscription	BL Navigation Paid Subscription	BL Navigation Subscribed	Purchased Windshield Wipers	Purchased In Cabin Air Filters		Remove Variable
Total Dollars Spent on Wiper Blades	0.1	1.0	1.0																					⚑	1
Total Dollars Spent on Cabin Air Filters	0.2	0.2	0.2	1.0	1.0																			⚑	1
Vehicle Age in Days	0.1	0.1	0.1	0.2	0.2	1.0	1.0																	⚑	1
Service Link Usage	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0														⚑	1
Average Miles Driven Per Month	0.9	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	1.0													⚑	1
Enrolled in Hyundai Blue Link (BL) Infotainment System	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0	0.0	1.0												🚩	2
BL Basic Service Subscribed	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0	0.0	1.0	0.5	0.5	1.0									🚩	3
BL Remote Start Subscribed	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.7	0.6	0.6	0.0	0.6	-0.1	0.8	0.6	0.4	0.8	1.0						⚑	1
BL Navigation Paid Subscription	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.4	0.5	0.0	0.5	-0.2	0.7	0.5	-0.1	0.9	0.8	-0.2	1.0				⚑	1
BL Navigation Subscribed	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.7	0.6	0.6	0.0	0.6	-0.1	0.7	0.6	0.4	0.8	1.0	0.5	0.8	1.0			🚩	2
Purchased Windshield Wipers	0.1	0.9	0.9	0.2	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.1	1.0		🚩	2
Purchased In Cabin Air Filters	0.2	0.2	0.2	0.9	0.9	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.2	1.0	🚩	2

Preparation of the final data set utilized the following variable transformation tactics. Some variables are derived, such as Average Customer Paid Dollars Per Repair Order and Average Months Between Service Visits. Categorical variables, such as vehicle model, are converted into binary variables. Aggregated models with less than 5% product mix into an "Other" model category and removed reference variables like Sonata from model categories and Oil Filter from parts quantity purchased. Lastly, the three main customer satisfaction survey scores have been normalized to 5-point scale to have a consistent score range between 1 and 5.

Section 5: Data mining

The final cleaned data set of 4,072 records, 1 outcome and 41 predictors was randomly partitioned into 60% training and 40% validation data for each of the following classification model evaluations.

kNN Models: There were 4 kNN models comparing all variables vs excluding binary variables, and normalized vs not normalized data.

kNN Results

	kNN Model 1	kNN Model 2	kNN Model 3	kNN Model 4
Model Size	4,072 Total 2,443 Training 1,629 Validation	4,072 Total 2,443 Training 1,629 Validation	4,072 Total 2,443 Training 1,629 Validation	4,072 Total 2,443 Training 1,629 Validation
Option	Not Normalized	Not Normalized	Normalized	Normalized
Variables	15 Continuous 8 Ordinal & 18 Binary = 41 Total	15 Continuous & 8 Ordinal	15 Continuous 8 Ordinal & 18 Binary = 41 Total	15 Continuous & 8 Ordinal
Best K	17	17	20	20
Misclassification Rate	25.59853%	25.59853%	36.27993%	30.75506%
Sensitivity	0.701564	0.701564	0.625752	0.671480
Specificity	0.788221	0.788221	0.649123	0.714286
ROC Curve AUC	0.811655	0.811655	0.685411	0.745463

kNN Confusion Matrix

kNN Model 1

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	583	248
0	169	629

kNN Model 2

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	583	248
0	169	629

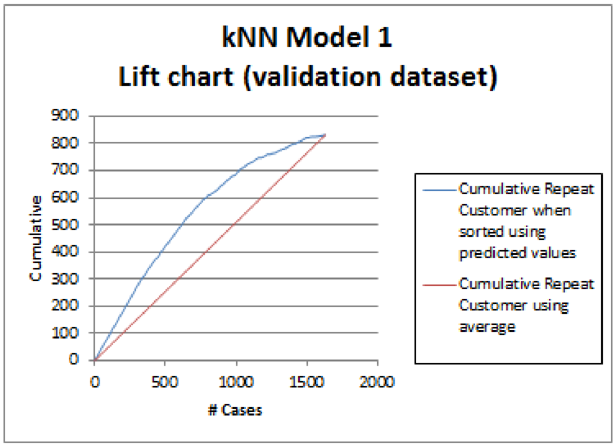
kNN Model 3

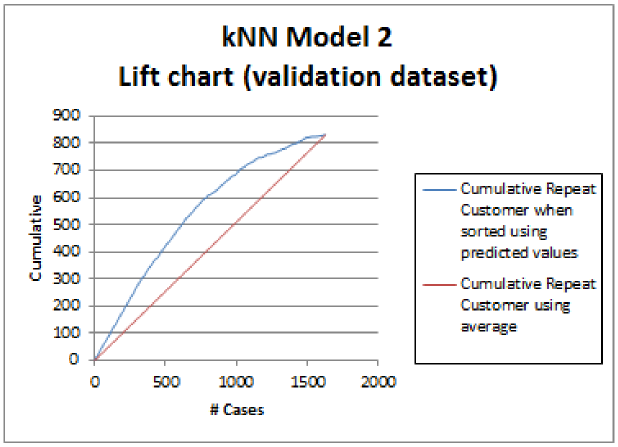
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	520	311
0	280	518

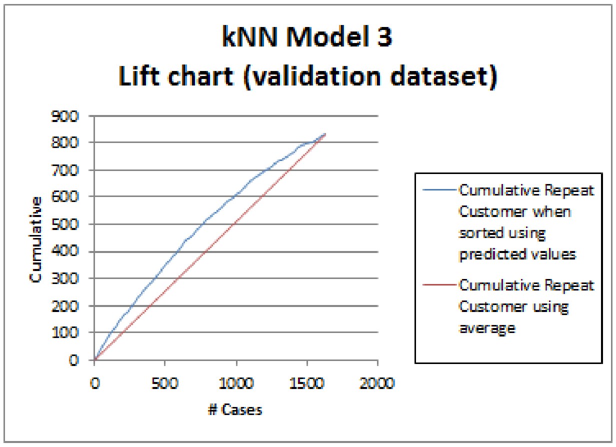
kNN Model 4

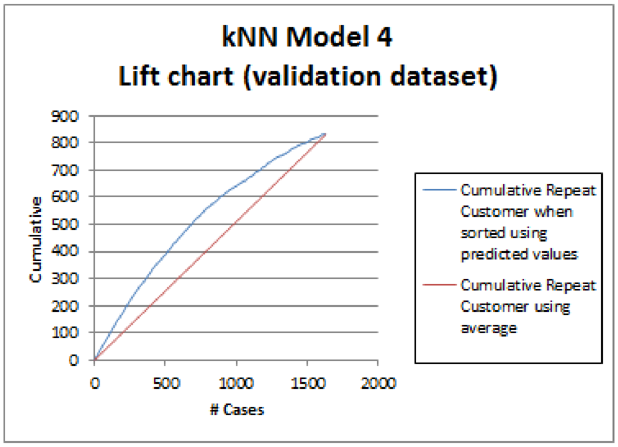
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	558	273
0	228	570

kNN Lift Chart









Logistic Regression Models:

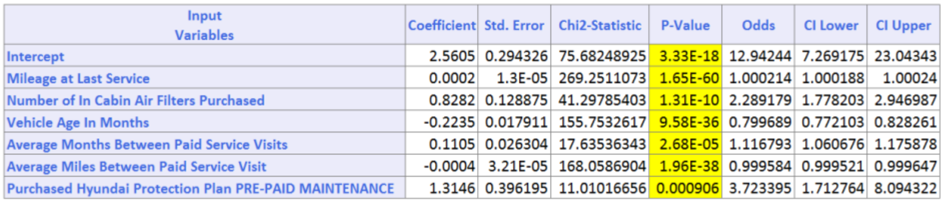
Log Results

	Log Model 1	**Log Model 2**	**Log Model 3**
Model Size	4,072 Total	4,072 Total	4,072 Total

	2,443 Training 1,629 Validation	2,443 Training 1,629 Validation	2,443 Training 1,629 Validation
Variable Selection	Backward	Forward	Stepwise
Cutoff	0.5	0.5	0.5
Number of Input Variables in Subset	7 Total: 1 Intercept 5 Continuous 1 Binary	6 Total: 1 Intercept 5 Continuous	6 Total: 1 Intercept 5 Continuous
Misclassification Rate	23.14303%	22.95887%	22.95887%
Sensitivity	0.719615	0.718412	0.718412
Specificity	0.819549	0.824561	0.824561
ROC Curve AUC	0.846087	0.846022	0.846022

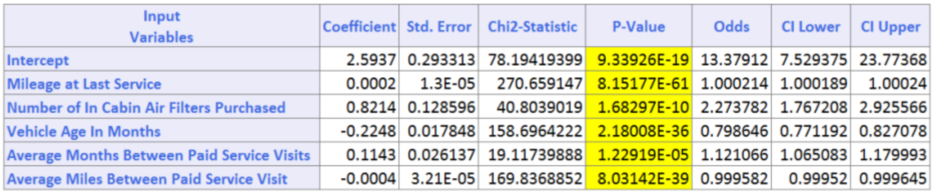
Log Regression Model

Regression Model 1



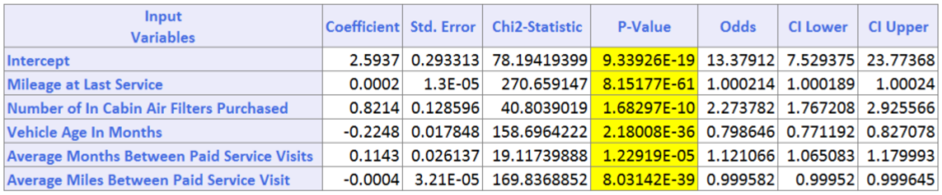
Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	2.5605	0.294326	75.68248925	3.33E-18	12.94244	7.269175	23.04343
Mileage at Last Service	0.0002	1.3E-05	269.2511073	1.65E-60	1.000214	1.000188	1.00024
Number of In Cabin Air Filters Purchased	0.8282	0.128875	41.29785403	1.31E-10	2.289179	1.778203	2.946987
Vehicle Age In Months	-0.2235	0.017911	155.7532617	9.58E-36	0.799689	0.772103	0.828261
Average Months Between Paid Service Visits	0.1105	0.026304	17.63536343	2.68E-05	1.116793	1.060676	1.175878
Average Miles Between Paid Service Visit	-0.0004	3.21E-05	168.0586904	1.96E-38	0.999584	0.999521	0.999647
Purchased Hyundai Protection Plan PRE-PAID MAINTENANCE	1.3146	0.396195	11.01016656	0.000906	3.723395	1.712764	8.094322

Regression Model 2



Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	2.5937	0.293313	78.19419399	9.33926E-19	13.37912	7.529375	23.77368
Mileage at Last Service	0.0002	1.3E-05	270.659147	8.15177E-61	1.000214	1.000189	1.00024
Number of In Cabin Air Filters Purchased	0.8214	0.128596	40.8039019	1.68297E-10	2.273782	1.767208	2.925566
Vehicle Age In Months	-0.2248	0.017848	158.6964222	2.18008E-36	0.798646	0.771192	0.827078
Average Months Between Paid Service Visits	0.1143	0.026137	19.11739888	1.22919E-05	1.121066	1.065083	1.179993
Average Miles Between Paid Service Visit	-0.0004	3.21E-05	169.8368852	8.03142E-39	0.999582	0.99952	0.999645

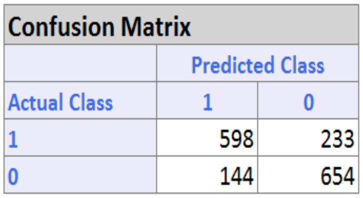
Regression Model 3



Input Variables	Coefficient	Std. Error	Chi2-Statistic	P-Value	Odds	CI Lower	CI Upper
Intercept	2.5937	0.293313	78.19419399	9.33926E-19	13.37912	7.529375	23.77368
Mileage at Last Service	0.0002	1.3E-05	270.659147	8.15177E-61	1.000214	1.000189	1.00024
Number of In Cabin Air Filters Purchased	0.8214	0.128596	40.8039019	1.68297E-10	2.273782	1.767208	2.925566
Vehicle Age In Months	-0.2248	0.017848	158.6964222	2.18008E-36	0.798646	0.771192	0.827078
Average Months Between Paid Service Visits	0.1143	0.026137	19.11739888	1.22919E-05	1.121066	1.065083	1.179993
Average Miles Between Paid Service Visit	-0.0004	3.21E-05	169.8368852	8.03142E-39	0.999582	0.99952	0.999645

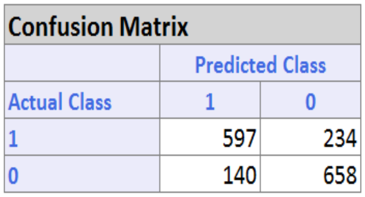
Log Confusion Matrix

Model 1



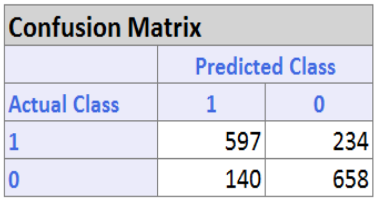
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	598	233
0	144	654

Model 2



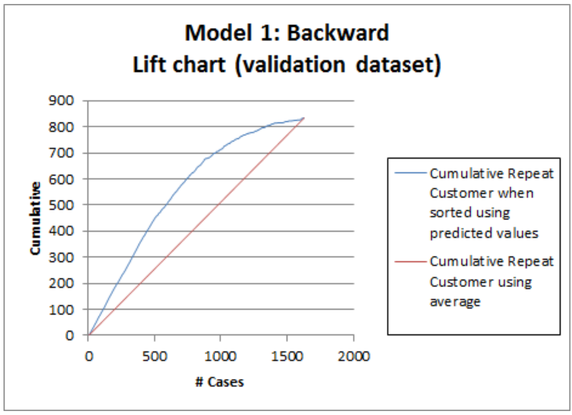
Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	597	234
0	140	658

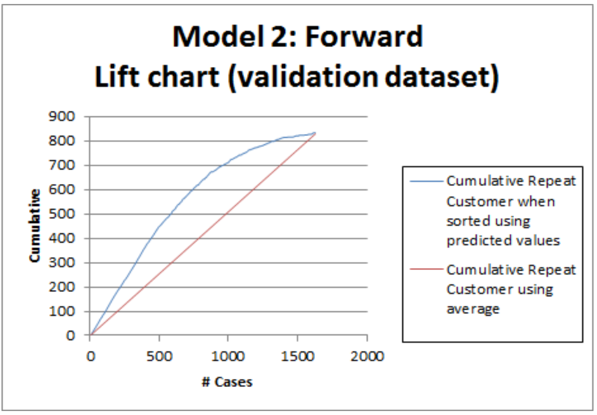
Model 3

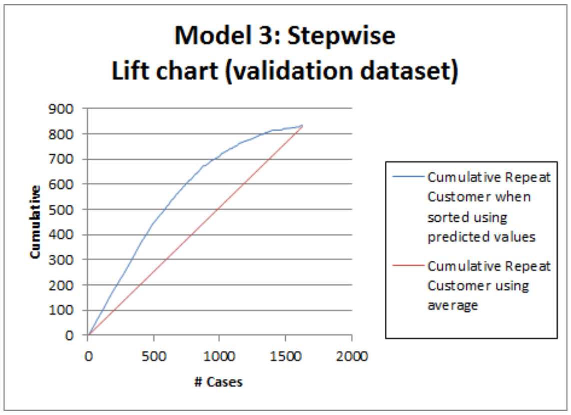


Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	597	234
0	140	658

Log Lift Chart





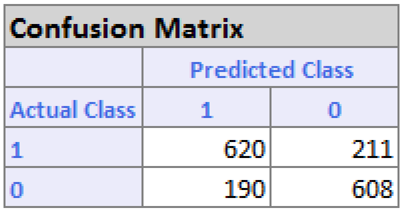


CART Model:

CART Results

	Classification Tree Model
Model Size	4,072 Total 2,443 Training 1,629 Validation
Variables	15 Continuous, 8 Ordinal & 18 Binary = 41 Total
Misclassification Rate	24.6163%
Sensitivity	0.746089
Specificity	0.761905
ROC Curve AUC	0.782606

CART Confusion Matrix



Confusion Matrix

	Predicted Class	
Actual Class	1	0
1	620	211
0	190	608

CART Lift Chart

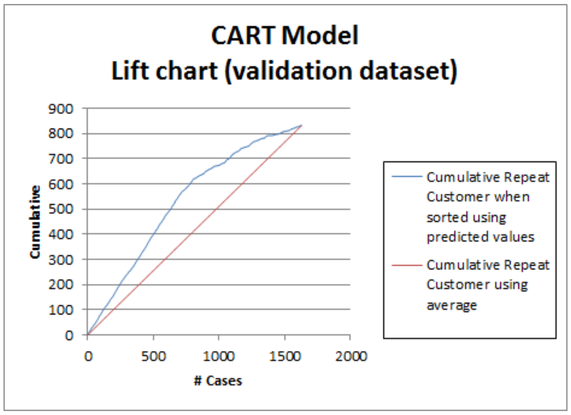


Table 2: Comparison of Model Performance

Best Results for each Model	**kNN Model**	**Logistic Model**	**Classification Tree Model**
Model Size	4,072 Total 2,443 Training 1,629 Validation	4,072 Total 2,443 Training 1,629 Validation	4,072 Total 2,443 Training 1,629 Validation
Variables	15 Continuous & 8 Ordinal = 23 Total	5 Continuous	15 Continuous, 8 Ordinal & 18 Binary = 41 Total
Misclassification Rate	25.59853%	22.95887%	24.6163%
Sensitivity	0.701564	0.718412	0.746089
Specificity	0.788221	0.824561	0.761905
ROC Curve AUC	0.811655	0.846022	0.782606

Results & Interpretation

a. kNN

Four versions of the kNN model were used to experiment on different dimensions: inclusion/exclusion of binary variables, and normalization of the data. Model 1 and 2 did not use normalized data, while Models 3 and 4 did. Models 1 and 3 included binary variables, while Models 2 and 4 excluded them.

It was found that the binary variables only contribute noise. This becomes evident when comparing the results of kNN Model 1 and kNN Model 2. Model 1 includes all variables while Model 2 excludes the binary variables, yet they yield the same results. It can then be concluded that the binary variables do not contribute to the core structure of the data. kNN Models 3 and 4 mirror this trend; Model 3 includes all variables and Model 4 excludes the binary variables, with Model 3 demonstrating worse results compared to 4.

It was also found that normalizing the data has a negative impact on the results. Normalizing the data (particularly the continuous variables) causes the data to lose its core structure, fuzzing the results to have a higher misclassification. Both kNN model 3 and kNN model 4 have higher Misclassification Rates than their non-normalized versions, kNN model 1 and kNN model 2.

Ultimately, kNN model 2 is determined to be the best of the 4 kNN models, where only the 18 continuous and 8 ordinal variables were included and the data was not normalized. This model is more parsimonious than kNN model 1 and also yields the same results.

b. Logistic Regression

Three logistic regression models were built using different variable selection methods: backward, forward, and stepwise. XLMiner automatically stopped the variable selection process and chose a variable subset with lowest value of complexity to represent the model. With a 0.5 cutoff value, these three models have misclassification rates of 23.14303%, 22.95887%, and 22.95887%, respectively.

Although model 1 contains more input variables than model 2 and model 3, it has the highest error rates among the models. This indicates that the binary variable “Purchased Hyundai Protection Plan PRE-PAID MAINTENANCE” does not contribute in improving the model.In addition, all of the three models have the same sensitivity of 0.72, which suggests that these models have the same capability in identifying the important class of “Repeat Customer”. Thus, the model that has the minimum misclassification rate should be chosen as the best model based on the performance evaluation rules. That is to say, either the model with forward selection or the model with stepwise selection should be the best model among the three versions.

The logit equation for both model 2 and model 3 is:

$$\mathit{Logit} = 2.5937 + 0.0002\mathit{MileageAtLastService} + 0.8214\mathit{NumberofInCabinAirFiltersPurchased} - 0.2248\mathit{VehicleAgeInMonths} + 0.1143\mathit{AverageMonthsBetweenPaidServiceVisits} - 0.0004\mathit{AverageMilesBetweenPaidServiceVisit}$$

The equation shows that customers whose vehicles have higher mileage at last service, more number of in-cabin filters purchased or longer time between paid service visits tend more to return to Hyundai dealership for maintenance services, whereas customers whose vehicles are older or have higher average miles between paid service visits are less likely to return.

c. Classification Tree

Aside from the Repeat Customer field, the Classification Tree model selected all variables from the data set as inputs. For the parsimony of the model, the best pruned tree model will be used for classification tree model, which produced 14 decision nodes in the model.

The best pruned tree model can be seen at the appendix.

A customer is classified based on the following seven decision variables:

1. Customer Paid RO Dollars
2. Mileage at Last Service
3. Average Months Between Paid Service Visit
4. Vehicle Age in Months
5. Average Miles Between Paid Service Visit
6. Average Customer Paid Dollars per RO
7. Last Service Dealer's Service Satisfaction Score

To summarize the model, we need to understand the major tree leaves and splits of the model. Customer Paid RO Dollars, Mileages at Last Service and Average Months Between Paid Service Visit are the major splits contain more than 100 observations on both sides of the nodes.

Three major leaves had same outcome of 0, which is not a repeat customer, that contains over 100 observations and a leaf of repeat customer had 623 observation. The largest leaf with outcome of 0 had 319 observations. Customers are less likely to revisit in such cases, 319 observations spend less than $113 and had their last service at 8,120 miles or less. Alternatively, customers are more likely to be a repeat customer (outcome of 1) when they spend more than $113 for their services, have service every 5 months or less and had their vehicle last service at 7149 miles or more.

The misclassification rate, sensitivity, specificity, ROC curve AUC are all listed on Table 2 in the previous section. The performance of our classification tree model is relatively close to our other models which are kNN and logistic model. To differentiate the differences from our other models, the Model's sensitivity and the specificity recorded close results by having 74.6089% and 76.1905% respectively. With that being said, the error rate for both classes are more balanced in this tree model.

Section 7: Conclusions

Comparing the models to each other, Logistic Regression outperforms kNN and Classification Tree on most key measures. The model has the lowest misclassification rate, as well as the highest Specificity and ROC Curve AUC. The Sensitivity for the Logistic Regression models falls between the kNN and Classification Tree models.

With the Logistic Regression model chosen, recommendations can be made based upon the results of the model. For instance, it is shown that customers with older vehicles and higher average miles between service visits are the least likely customers to visit the dealership. These factors indicate that little to no relationship exists between these customers and the dealership, due to less coverage for older vehicles or having established no initial loyalty after the initial purchase. Hyundai may find it useful to email service reminders with more attractive coupons and offers to customers that have not visited a dealership within seven months, which may to entice them to visit again.

Meanwhile, it is in the interest of the dealership to act on customers that purchase many in-cabin filters, have a higher mileage recorded at their most recent service, and have a longer time between service visits. These associated factors are consistent with the profile of a customer that relies on the dealership for infrequent car maintenance tasks. With this knowledge, dealership management can push replacement air filters and related services to new and recent customers, because they are the most common drivers of repeat dealership visits.

Alternatively, management could consider utilizing one of the other models to develop strategies. The classification tree model shares the same variables as the Logistic Regression model, so it is possible for managers to draw conclusions and develop similar programs for the dealership. Rather than focusing on the factors identified by the Logistic model for push marketing aimed at the newest customers, Hyundai can use CART decision points to inform and deploy a pull campaign.This strategy seeks to trigger customer action processes and hopefully recapture potential service defectors.

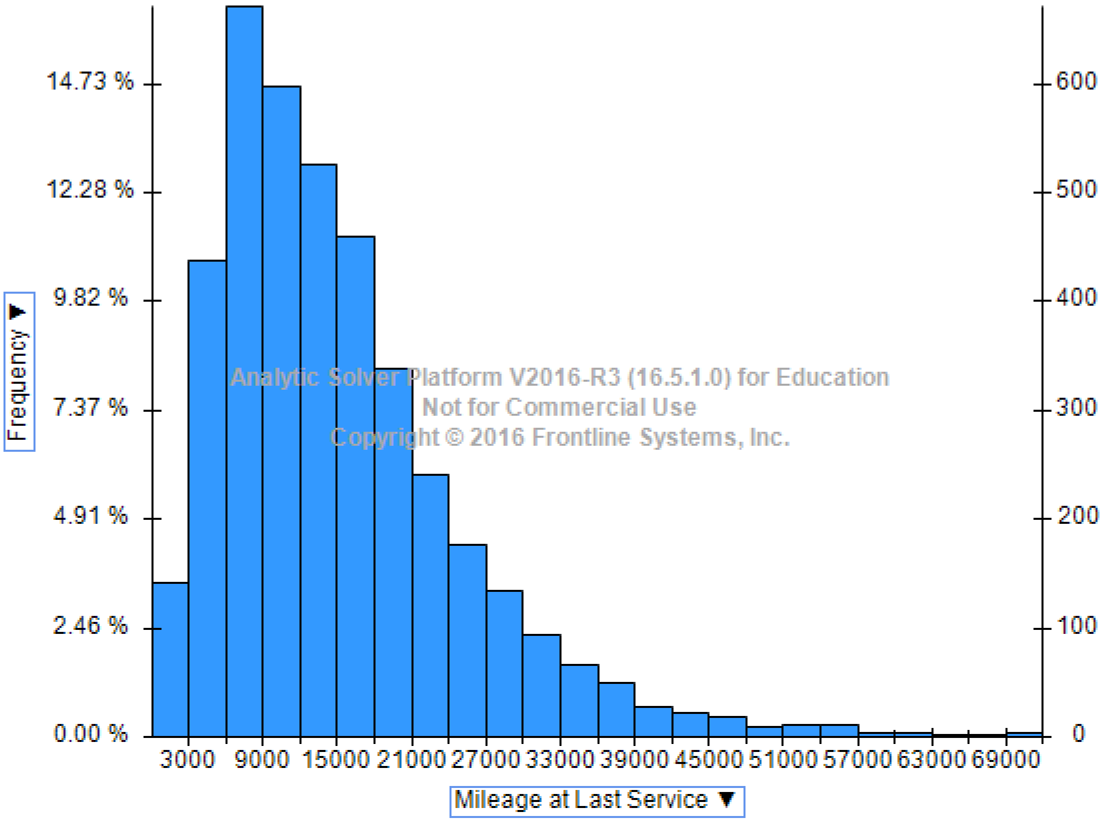
Upon reflection, the investigation was an interesting introduction to data mining through the use of predictive models, but there were difficulties. The record limitations and thresholds of XLMiner have been noted, and R is not easy or intuitive for first time users. Replicating the study with alternative software such as SAS or Alteryx may accomplish the same tasks with greater ease and effectiveness.

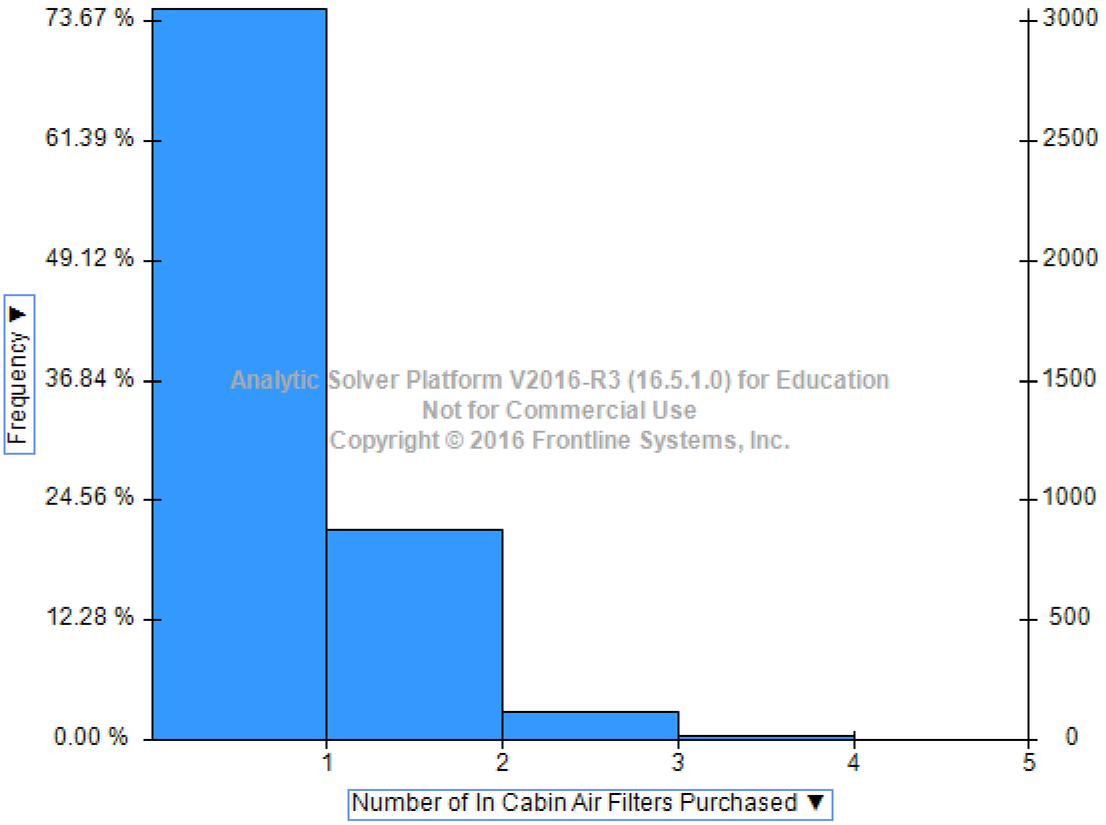
The results were straightforward and expected, but it would have been nice to see if Hyundai’s various service retention programs (e.g., Blue Link, Rewards and HPP Prepaid Maintenance) and customer satisfaction initiatives (e.g., decreasing prices and increasing customer satisfaction) were making a greater impact. In the future, managers will want to investigate the effectiveness of these programs, so data must be structured in a way that facilitates such investigation.

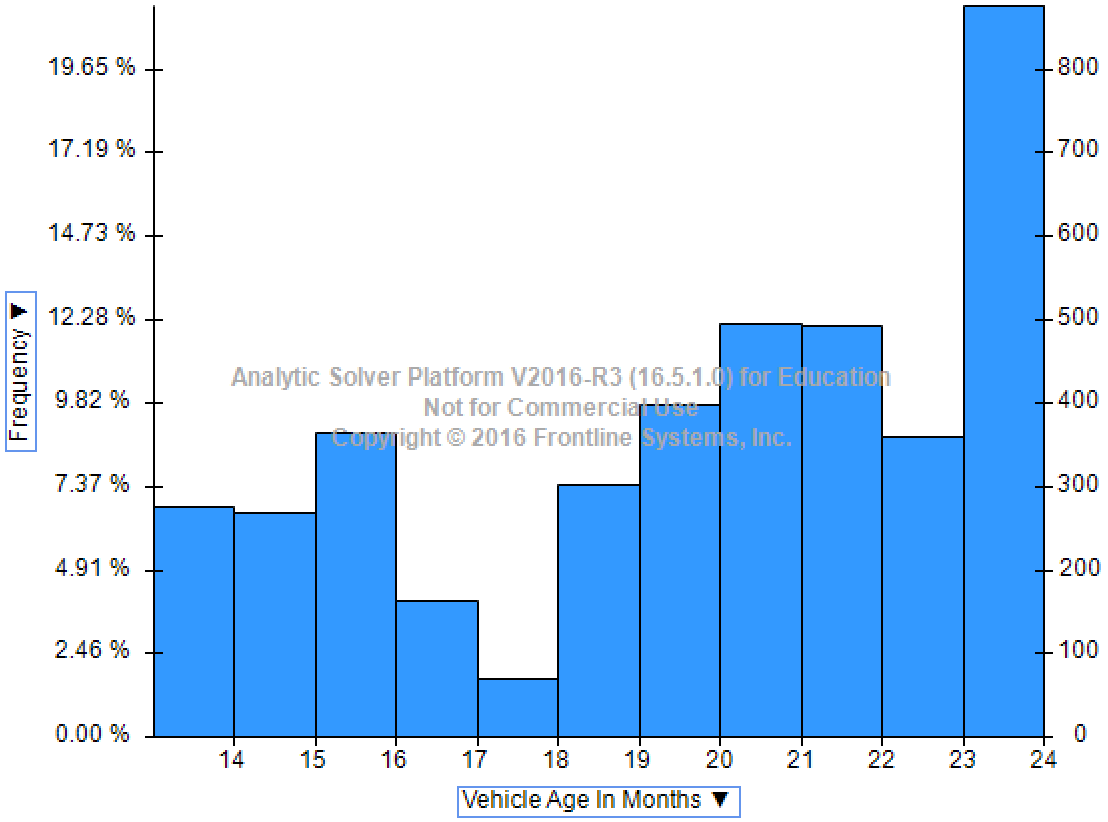
Other considerations for future investigations might include performing data mining on additional historical customer data or building models for older 3-year+ owners, as the current study only samples 1st and 2nd year customers from the year 2016. The study could be expanded using data from more previous years, which may capture trends that were not found in the 2016 data.

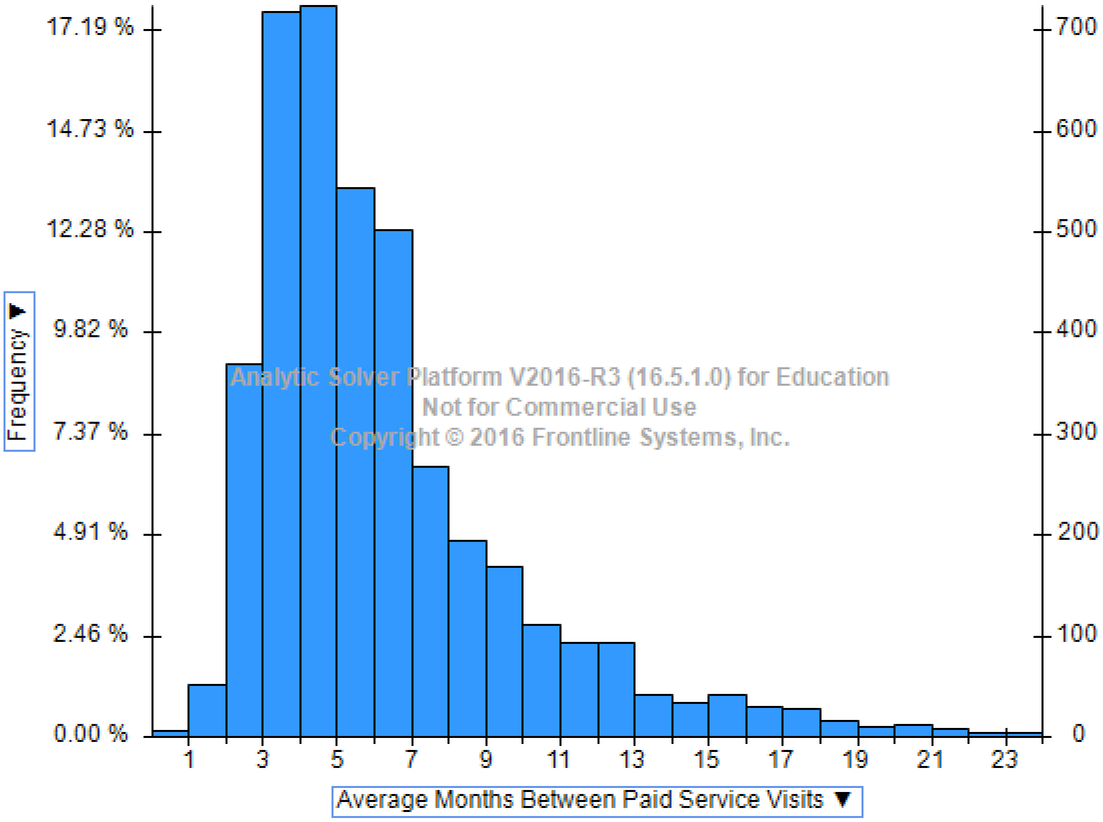
Appendix

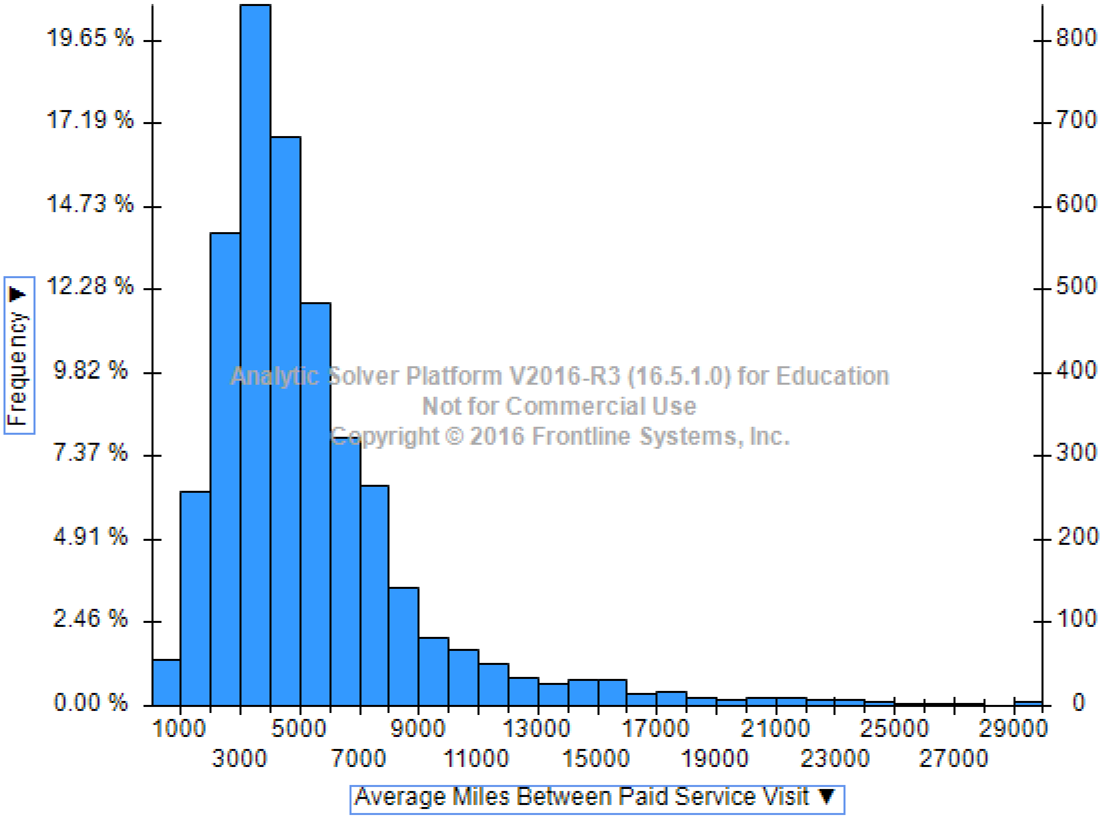
Histograms:



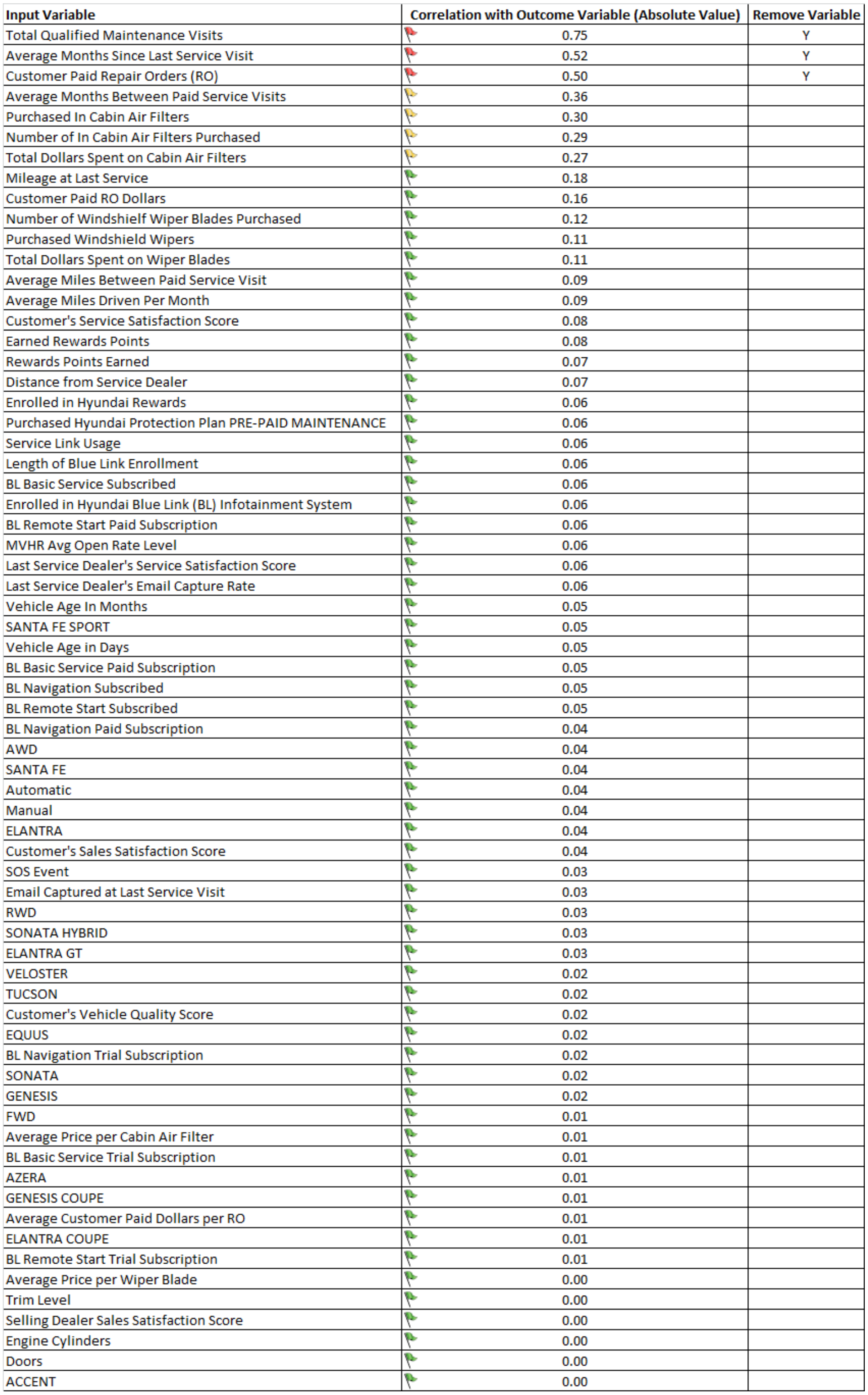

73.67 %
61.39 %
49.12 %
36.84 %
24.56 %
12.28 %
0.00 %
Frequency ▼
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
Number of In Cabin Air Filters Purchased ▼


19.65 %
17.19 %
14.73 %
12.28 %
9.82 %
7.37 %
4.91 %
2.46 %
0.00 %
Frequency ▼
800
700
600
500
400
300
200
100
0
14
15
16
17
18
19
20
21
22
23
24
Vehicle Age In Months ▼
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.


17.19 %
14.73 %
12.28 %
9.82 %
7.37 %
4.91 %
2.46 %
0.00 %
Frequency ▼
700
600
500
400
300
200
100
0
1
3
5
7
9
11
13
15
17
19
21
23
Average Months Between Paid Service Visits ▼
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

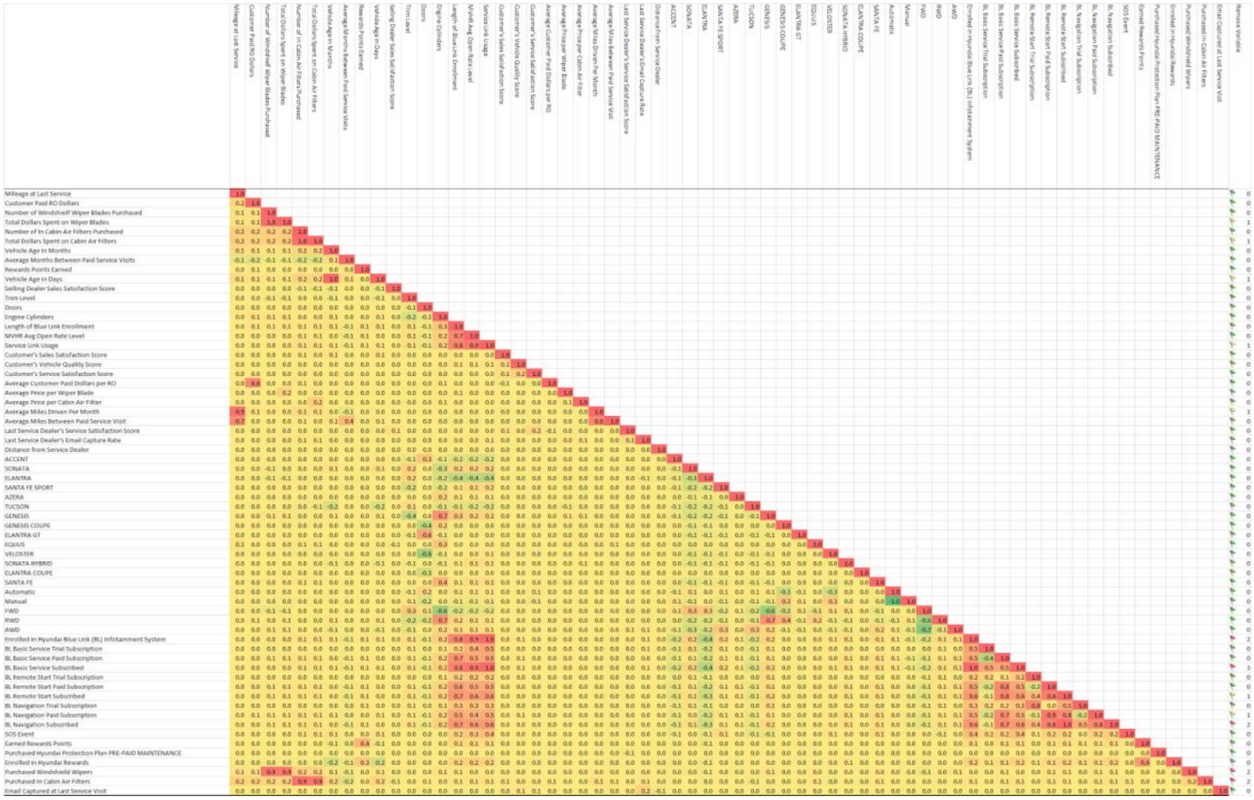

Frequency ▼
19.65 %
17.19 %
14.73 %
12.28 %
9.82 %
7.37 %
4.91 %
2.46 %
0.00 %
800
700
600
500
400
300
200
100
0
1000
3000
5000
7000
9000
11000
13000
15000
17000
19000
21000
23000
25000
27000
29000
Average Miles Between Paid Service Visit ▼
Analytic Solver Platform V2016-R3 (16.5.1.0) for Education
Not for Commercial Use
Copyright © 2016 Frontline Systems, Inc.

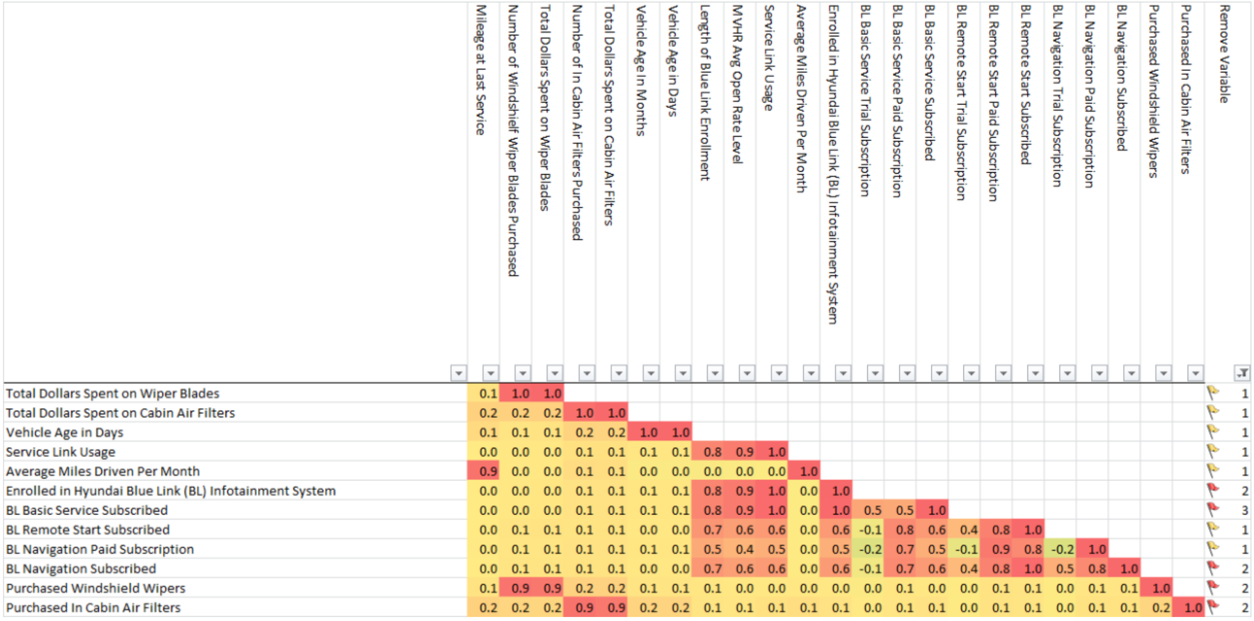
Correlation Analysis between Input and Output Variables:



Input Variable	Correlation with Outcome Variable (Absolute Value)	Remove Variable
Total Qualified Maintenance Visits	🟥 0.75	Y
Average Months Since Last Service Visit	🟥 0.52	Y
Customer Paid Repair Orders (RO)	🟥 0.50	Y
Average Months Between Paid Service Visits	🟨 0.36	
Purchased In Cabin Air Filters	🟨 0.30	
Number of In Cabin Air Filters Purchased	🟨 0.29	
Total Dollars Spent on Cabin Air Filters	🟨 0.27	
Mileage at Last Service	🟩 0.18	
Customer Paid RO Dollars	🟩 0.16	
Number of Windshielf Wiper Blades Purchased	🟩 0.12	
Purchased Windshield Wipers	🟩 0.11	
Total Dollars Spent on Wiper Blades	🟩 0.11	
Average Miles Between Paid Service Visit	🟩 0.09	
Average Miles Driven Per Month	🟩 0.09	
Customer's Service Satisfaction Score	🟩 0.08	
Earned Rewards Points	🟩 0.08	
Rewards Points Earned	🟩 0.07	
Distance from Service Dealer	🟩 0.07	
Enrolled in Hyundai Rewards	🟩 0.06	
Purchased Hyundai Protection Plan PRE-PAID MAINTENANCE	🟩 0.06	
Service Link Usage	🟩 0.06	
Length of Blue Link Enrollment	🟩 0.06	
BL Basic Service Subscribed	🟩 0.06	
Enrolled in Hyundai Blue Link (BL) Infotainment System	🟩 0.06	
BL Remote Start Paid Subscription	🟩 0.06	
MVHR Avg Open Rate Level	🟩 0.06	
Last Service Dealer's Service Satisfaction Score	🟩 0.06	
Last Service Dealer's Email Capture Rate	🟩 0.06	
Vehicle Age In Months	🟩 0.05	
SANTA FE SPORT	🟩 0.05	
Vehicle Age in Days	🟩 0.05	
BL Basic Service Paid Subscription	🟩 0.05	
BL Navigation Subscribed	🟩 0.05	
BL Remote Start Subscribed	🟩 0.05	
BL Navigation Paid Subscription	🟩 0.04	
AWD	🟩 0.04	
SANTA FE	🟩 0.04	
Automatic	🟩 0.04	
Manual	🟩 0.04	
ELANTRA	🟩 0.04	
Customer's Sales Satisfaction Score	🟩 0.04	
SOS Event	🟩 0.03	
Email Captured at Last Service Visit	🟩 0.03	
RWD	🟩 0.03	
SONATA HYBRID	🟩 0.03	
ELANTRA GT	🟩 0.03	
VELOSTER	🟩 0.02	
TUCSON	🟩 0.02	
Customer's Vehicle Quality Score	🟩 0.02	
EQUUS	🟩 0.02	
BL Navigation Trial Subscription	🟩 0.02	
SONATA	🟩 0.02	
GENESIS	🟩 0.02	
FWD	🟩 0.01	
Average Price per Cabin Air Filter	🟩 0.01	
BL Basic Service Trial Subscription	🟩 0.01	
AZERA	🟩 0.01	
GENESIS COUPE	🟩 0.01	
Average Customer Paid Dollars per RO	🟩 0.01	
ELANTRA COUPE	🟩 0.01	
BL Remote Start Trial Subscription	🟩 0.01	
Average Price per Wiper Blade	🟩 0.00	
Trim Level	🟩 0.00	
Selling Dealer Sales Satisfaction Score	🟩 0.00	
Engine Cylinders	🟩 0.00	
Doors	🟩 0.00	
ACCENT	🟩 0.00	

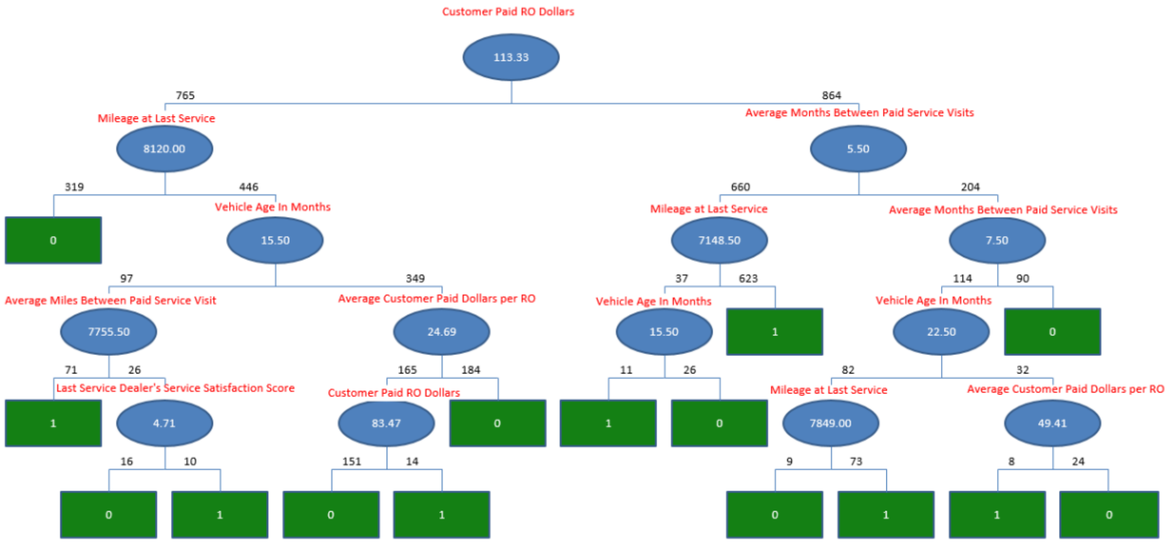
Correlation Matrix of Input Variables:





	Mileage at Last Service	Number of Windshielf Wiper Blades Purchased	Total Dollars Spent on Wiper Blades	Number of In Cabin Air Filters Purchased	Total Dollars Spent on Cabin Air Filters	Vehicle Age In Months	Vehicle Age in Days	Length of Blue Link Enrollment	MVHR Avg Open Rate Level	Service Link Usage	Average Miles Driven Per Month	Enrolled in Hyundai Blue Link (BL) Infotainment System	BL Basic Service Trial Subscription	BL Basic Service Paid Subscription	BL Basic Service Subscribed	BL Remote Start Trial Subscription	BL Remote Start Paid Subscription	BL Remote Start Subscribed	BL Navigation Trial Subscription	BL Navigation Paid Subscription	BL Navigation Subscribed	Purchased Windshield Wipers	Purchased In Cabin Air Filters	Remove Variable
Total Dollars Spent on Wiper Blades	0.1	1.0	1.0																					⚑ 1
Total Dollars Spent on Cabin Air Filters	0.2	0.2	0.2	1.0	1.0																			⚑ 1
Vehicle Age in Days	0.1	0.1	0.1	0.2	0.2	1.0	1.0																	⚑ 1
Service Link Usage	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0														⚑ 1
Average Miles Driven Per Month	0.9	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	1.0													⚑ 1
Enrolled in Hyundai Blue Link (BL) Infotainment System	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0	0.0	1.0												⚑ 2
BL Basic Service Subscribed	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.8	0.9	1.0	0.0	1.0	0.5	0.5	1.0									⚑ 3
BL Remote Start Subscribed	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.7	0.6	0.6	0.0	0.6	-0.1	0.8	0.6	0.4	0.8	1.0						⚑ 1
BL Navigation Paid Subscription	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.5	0.4	0.5	0.0	0.5	-0.2	0.7	0.5	-0.1	0.9	0.8	-0.2	1.0				⚑ 1
BL Navigation Subscribed	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.7	0.6	0.6	0.0	0.6	-0.1	0.7	0.6	0.4	0.8	1.0	0.5	0.8	1.0			⚑ 2
Purchased Windshield Wipers	0.1	0.9	0.9	0.2	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.1	1.0		⚑ 2
Purchased In Cabin Air Filters	0.2	0.2	0.2	0.9	0.9	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.0	0.1	0.1	0.2	1.0	⚑ 2

Classification Tree-Best Pruned Tree


Customer Paid RO Dollars
113.33
765
864
Mileage at Last Service
8120.00
Average Months Between Paid Service Visits
5.50
319
446
660
204
0
Vehicle Age In Months
15.50
Mileage at Last Service
7148.50
Average Months Between Paid Service Visits
7.50
97
349
37
623
114
90
Average Miles Between Paid Service Visit
7755.50
Average Customer Paid Dollars per RO
24.69
Vehicle Age In Months
15.50
1
Vehicle Age In Months
22.50
0
71
26
165
184
11
26
82
32
1
Last Service Dealer's Service Satisfaction Score
4.71
Customer Paid RO Dollars
83.47
0
1
0
Mileage at Last Service
7849.00
Average Customer Paid Dollars per RO
49.41
16
10
151
14
9
73
8
24
0
1
0
1
0
1
1
0