Predictive Analytics in Customer Retention Management (CRM) of Automotive Industry Using the Hadoop Ecosystem and SAS

Jun Long, junlong2015@u.northwestern.edu

Abstract— **Automotive industry has tremendous opportunities to leverage the Hadoop ecosystem to provide data storage and analysis. The Customer Relationship Management is an important part to assist in customer retention and drive sales growth. This study uses the Hadoop ecosystem to process a subset of the data to examine a real-world problem in the automotive CRM field. The study also uses SAS to predict probability of escalation when customers have problem with their vehicles. Logistic regression technique is used to build this predictive model. Key predictors driving customer retention are detected based on real world data gathered from automotive industry. The results can be applied to facilitate decision making and therefore serve better customer retention management.**

Index Terms— **Hadoop, Hive, Predictive Modeling, SAS, Customer Relationship Management, Automotive**

I. Introduction

This project highlights big data concepts and utilizes the Hadoop ecosystem as distributed by Cloudera plus SAS to describe and analyze the real-world data from company A. The automotive industry faces more competition nowadays due to product homogeneity. The heated competition provides consumers with more choices, which results in the increased customer acquisition cost and decreased customer loyalty. According to Pogol’s research in 2007, it costs at least 6-10 times more to acquire a new customer than it does to retain an existing one, and acquisition costs are a large portion of an automotive company's administrative and marketing cost. Therefore, customer retention becomes one of the key factors in sustaining business growth. Big data analytics and predictive modeling can be an optimal tool to serve this purpose.

Hadoop Ecosystem

The Hadoop Ecosystem Hadoop is an open source project that has developed and constantly upgrades and improves software to run on commodity hardware to store and process massive datasets. Because of its design it provides essentially unlimited scalability. Hadoop consists of two core components:

- The Hadoop Distributed File System (HDFS)
- MapReduce Software Framework

There are many other software systems based on and built around Hadoop which provide specific functionality which is often referred to as the Hadoop Ecosystem:

- Pig
- Hive
- Hue
- Hbase
- Flume
- Oozie
- Sqoop
- Zookeeper
- Others are being added on a regular basis.

The relationship between the various pieces of software is shown graphically below.

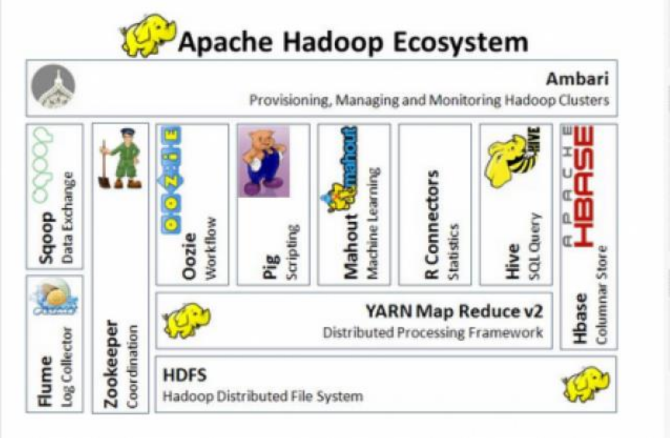


Figure 1: Hadoop Ecosystem Overview

The two key strengths of the Hadoop system are the fact that it is built on commodity hardware thus reducing the cost of expansion considerably and it is designed to expect failure of hardware components. The commodity hardware can be added in a modular fashion so the system can grow or shrink as needs dictate. Hadoop handles the anticipated failure of hardware components by building in massive redundancy in data.

The benefits of a “Big Data” Hadoop based solution over a traditional SQL/RDBMS solution are:

- The ability to quickly accommodate changing data schema and requirements including varied data formats such as media and document files in addition to data elements.
- The ability to easily scale the system up using

commodity hardware.

- Redundancy of data stored across multiple file systems.
- Reduce the upfront burden of developing an RDBMS schema that is not likely to change.

For this project the Hive query language was leveraged extensively for data processing and exploratory data analysis, then the data will be shipped to predictive modeling tools for advanced modeling.

Predictive Modeling and SAS

The most commonly used technologies in predictive modeling are regression (Linear, Logistic, etc.) and decision tree. In our case, both techniques are applied in our study, and the comparison results between them are also discussed in the results part.

SEMMA, one of the most well-known predictive modeling methodologies from SAS Institute Inc, is applied in this study. SEMMA lists the sequential steps which are Sample, Explore, Modify, Model and Assess. Once population of interest is defined, the next step would be gathering and cleansing data, then more explorations may be needed on the data including transformation and creating new variables. One of the key steps is to select samples for model development. It usually separates the dataset into two parts: one for modeling and the other for validation. Then the model can be built to identify significant predictor variables and model validation can be conducted with validation dataset.

SAS is developed by SAS Institute for statistical analysis and advanced analytics. It offers state-of-the-art predictive analytics and data mining capabilities that enable organizations to analyze complex data, find useful insights and act confidently to make fact-based decisions. Another advantage of SAS is that it enables quick comparison and results visualization. For this study we will use SAS to conduct modeling.

II. Modeling

Company A is a global automobile manufacturer, which designs, manufactures and distributes passenger and commercial vehicles, motorcycles, engines, and turbo-machinery, and offers related services including financing, leasing and fleet management. This company maintains the largest market share in some continents for over two decades.

Like all other automakers Company A has to deal with customer complain cases from all kinds of sources such as phone, letter, email, etc. Sometimes it has to handle high escalation cases from attorney demand letters, third parties (e.g. Better Business Bureau) or field representatives, which usually cause great losses. One most common way to deal with the cases is to offer goodwill, but for those high escalation cases Company A has to provide special resolutions including:

- Cash Settlement: Customer receives monetary payment, signs release and keeps vehicle.
- Trade Assistance: Customer buys new car at dealer invoice, company contributes to down payment, dealer keeps traded vehicle and resells it.
- Repurchase: Customer receives refund and leaves the brand, vehicle is sold at auction.
- Replacement: Customer receives new car (same sales contract), stays with brand, old vehicle is sold at auction.

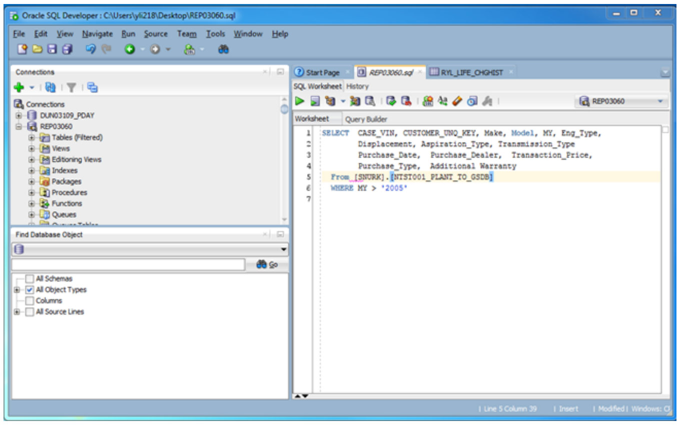
Modeling Purposes

The major purposes of modeling, from the introduction above, would answer questions like: which types of customers are more likely to escalate and what costly and burdensome resolutions are required by those customers (e.g. repurchase, replacement, trade assistance, or cash settlement)? How to determine the most appropriate response to return customer to satisfied state, and therefore, maintain brand loyalty?

Predictive modeling techniques can be used to answer these questions. More specific, our interest is to predict whether customers escalate or not, thus logistic regression and CART can be suitable for this case. The population can be defined as “hand-raisers” which means those customers who called for complaining over the past five years from all possible sources (phone, letter, e-mail, executive management, dealer, etc.). However, some cases should be excluded, because the model is to help Company A make decision when a customer calls for complaint. If customers escalated the case without even “raising their hands”, they are of less interests to the company since they are unpredictable. This model is named “Probability of Escalation (POE)”.

Data Collection and Potential Predictor Variables

The raw data in our study is provided by Company A’s IT team using Oracle (see figure 2), which is vehicle sales and cases information for hand-raisers.



```sql
SELECT  CASE_VIN, CUSTOMER_UNQ_KEY, Make, Model, MY, Eng_Type,
        Displacement, Aspiration_Type, Transmission_Type
        Purchase_Date,  Purchase_Dealer,  Transaction_Price,
        Purchase_Type,  Additional Warranty
  From [SNURK].[NTST001_PLANT_TO_GSDB]
  WHERE MY > '2005'
```

Figure 2: ETL Tool for Raw Data

The sales data goes back to 10 years ago, because vehicles usually have longer replacement cycle time. Using longer historical data can ensure enough ownership information is collected. The cases data only traces to 5 years ago according to company’s interest. When the raw data is received, we use Hive to run queries and join multiple tables based on customer ID (see figure 3). After cleansing the data, in total 110879 valid records were uniquely identified. Potential predictors are chosen from the following 4 categories:

Ownership History

- Total share of garage for the customer

Current share of garage for the customer

Purchase Behavior

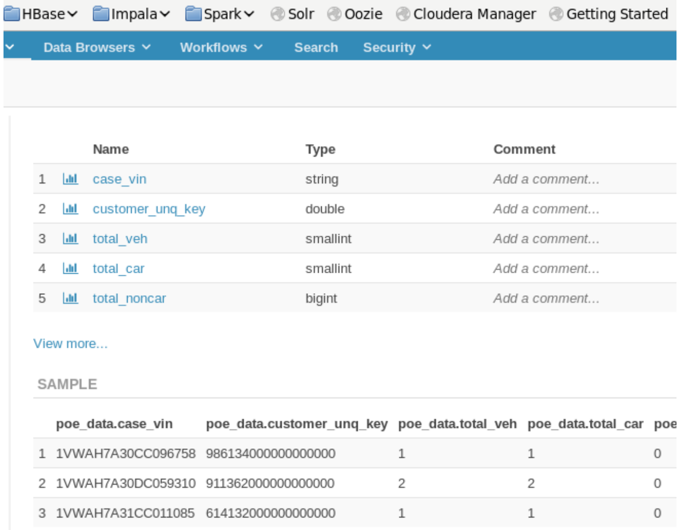
- Recency: most recent purchase from current date (months) for each customer
- Longevity: the very first purchase from current date (months) for each customer
- Mean time between purchases (MTBP) for each customer if more than one vehicle
- Number and percentage of additional warranty for each customer
- Number and percentage of new versus used vehicles for each customer
- Number and percentage of purchased versus leased vehicles for each customer

Claimed Vehicle Info

- First year of model or not for the claimed vehicle
- Number of claims for the claimed vehicle
- Number of severity index count for the claimed vehicle
- Maximum days down for the claimed vehicle
- Total number of cases for the claimed vehicle
- Case sources for the claimed vehicle
- Type of assistance requested by customer for the claimed vehicle
- Early failure indicator: time between purchase date and first case open date for the claimed vehicle, if less than 12 months it can be considered as early failure

Dealer Info

- Geographic region where the claimed vehicle belongs
- Dealer service rating (customer experience scores) where the claimed vehicle belongs



	Name	Type	Comment
1	case_vin	string	Add a comment...
2	customer_unq_key	double	Add a comment...
3	total_veh	smallint	Add a comment...
4	total_car	smallint	Add a comment...
5	total_noncar	bigint	Add a comment...

View more...

SAMPLE

	poe_data.case_vin	poe_data.customer_unq_key	poe_data.total_veh	poe_data.total_car	poe
1	1VWAH7A30CC096758	986134000000000000	1	1	0
2	1VWAH7A30DC059310	911362000000000000	2	2	0
3	1VWAH7A31CC011085	614132000000000000	1	1	0

Figure 3: Using Hive for Data Pre-processing

Other predictors such as customer and household demographics (gender, age, income, marital status, number of children, education level etc.) and quality rating for each car line, may also have impact on final probability of escalation, but unfortunately they are currently not available from Company A.

Result Analysis

After cleansing the dataset, decision tree and logistic regression models are run through SAS Enterprise Miner 7.1. Figure 3.1 below shows the process flow of SAS Enterprise Miner, we first use Partition Node to partition data into 70% for training and 30% for validation using stratified sampling on the target variable. User defined interactions and second-degree polynomial terms are added in Enterprise Miner Regression node. Stepwise selection with both entry level and stay level of 0.05 is used for final variable selection. Selection criteria are set to minimize the validation misclassification rate. At last, Comparison Node can compare the goodness of each model automatically by multiple criteria.

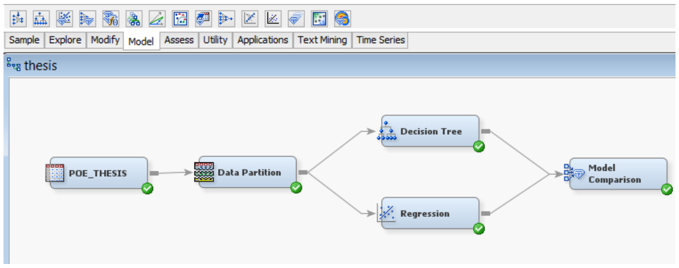
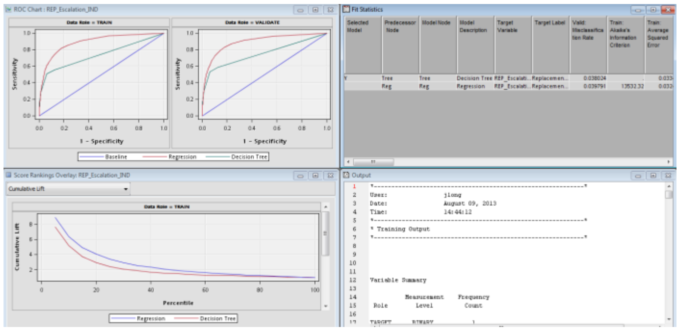


Figure 3.1: Process Flow in SAS Enterprise Miner

Which model works better in this case?

The results of model comparison between logistic regression and decision tree are shown in Figure 4 below. Note that Receiver Operating Characteristic (ROC) curve of Logistic Regression is much steeper than of that of Decision Tree. In other words, ROC of Logistic Regression is better than Decision Tree. From cumulative lift chart, we can draw the same conclusion that logistic regression model get better prediction in this case.



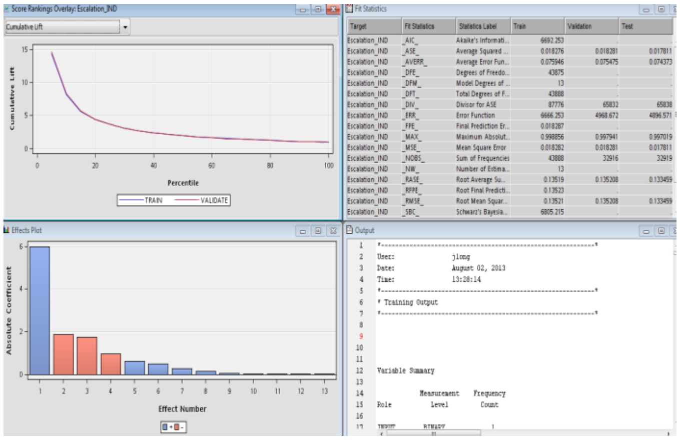
Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Tree	Tree	Decision Tree	REP_Escalati...	Replacemen...	0.038024	.	0.033
	Reg	Reg	Regression	REP_Escalati...	Replacemen...	0.039791	13532.32	0.032

```
*------------------------------------------------------------*
User:               jlong
Date:               August 09, 2013
Time:               14:44:12
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement    Frequency
 Role         Level         Count

TARGET      BINARY           1
```

Figure 4: Model Results Comparison

Significant Predictors

Figure 5 below shows the output of Logistic Regression, which includes four parts that are Cumulative Lift Chart, Fits Statistics, Effects Plot, and Output Summary Table. Effects Plot tells significant predictors and their relevant importance. Notice there are two colored effects: one is positive and the other is negative. Positive effects, which have positive coefficients, mean that variables increase Probability of Escalation (POE), and vice versa.

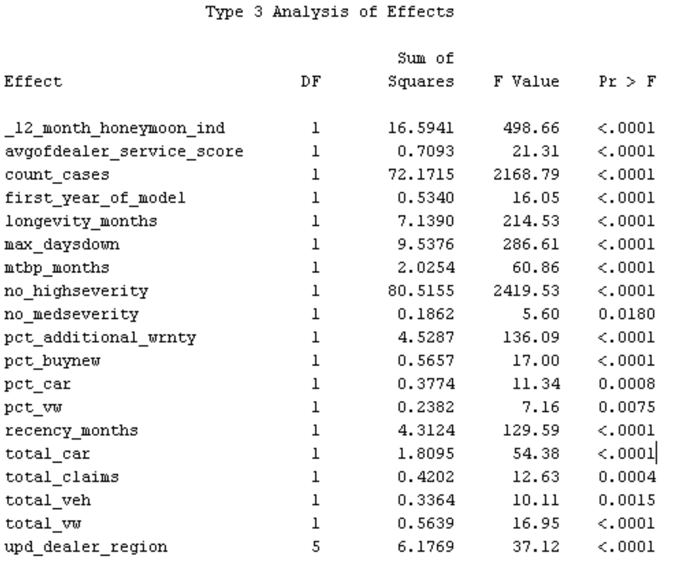


Target	Fit Statistics	Statistics Label	Train	Validation	Test
Escalation_IND	_AIC_	Akaike's Informati...	6692.253	.	.
Escalation_IND	_ASE_	Average Squared ...	0.018276	0.018281	0.017811
Escalation_IND	_AVERR_	Average Error Fun...	0.075946	0.075475	0.074373
Escalation_IND	_DFE_	Degrees of Freedo...	43875	.	.
Escalation_IND	_DFM_	Model Degrees of ...	13	.	.
Escalation_IND	_DFT_	Total Degrees of F...	43888	.	.
Escalation_IND	_DIV_	Divisor for ASE	87776	65832	65838
Escalation_IND	_ERR_	Error Function	6666.253	4968.672	4896.571
Escalation_IND	_FPE_	Final Prediction Er...	0.018287	.	.
Escalation_IND	_MAX_	Maximum Absolut...	0.998856	0.997941	0.997019
Escalation_IND	_MSE_	Mean Square Error	0.018282	0.018281	0.017811
Escalation_IND	_NOBS_	Sum of Frequencies	43888	32916	32919
Escalation_IND	_NW_	Number of Estima...	13	.	.
Escalation_IND	_RASE_	Root Average Su...	0.13519	0.135208	0.133459
Escalation_IND	_RFPE_	Root Final Predicti...	0.13523	.	.
Escalation_IND	_RMSE_	Root Mean Squar...	0.13521	0.135208	0.133459
Escalation_IND	_SBC_	Schwarz's Bayesia...	6805.215	.	.

```
*------------------------------------------------------------*
User:                 jlong
Date:                 August 02, 2013
Time:                 13:28:14
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary


             Measurement    Frequency
Role            Level         Count

INPUT         BINARY             1
```

Figure 5: Logistic Regression Output

A number of key factors that drive POE score are identified during the model building phase of the project. Figure 6 provides a better way to see the key drivers, which summarizes the F and P values of each significant predictor. The predictor variables that are used in final model may be slightly different than the variables showing in Figure 6. Because multi-collinearity needs to be taken care of, and some predictors need to be dropped.



```
                           Type 3 Analysis of Effects

                                            Sum of
Effect                         DF          Squares     F Value    Pr > F

_12_month_honeymoon_ind         1          16.5941      498.66    <.0001
avgofdealer_service_score       1           0.7093       21.31    <.0001
count_cases                     1          72.1715     2168.79    <.0001
first_year_of_model             1           0.5340       16.05    <.0001
longevity_months                1           7.1390      214.53    <.0001
max_daysdown                    1           9.5376      286.61    <.0001
mtbp_months                     1           2.0254       60.86    <.0001
no_highseverity                 1          80.5155     2419.53    <.0001
no_medseverity                  1           0.1862        5.60    0.0180
pct_additional_wrnty            1           4.5287      136.09    <.0001
pct_buynew                      1           0.5657       17.00    <.0001
pct_car                         1           0.3774       11.34    0.0008
pct_vw                          1           0.2382        7.16    0.0075
recency_months                  1           4.3124      129.59    <.0001
total_car                       1           1.8095       54.38    <.0001
total_claims                    1           0.4202       12.63    0.0004
total_veh                       1           0.3364       10.11    0.0015
total_vw                        1           0.5639       16.95    <.0001
upd_dealer_region               5           6.1769       37.12    <.0001
```

Figure 3.4: Analysis of Effects

III. Results Discussion

The cumulative gains chart of captured target and ROC curve generated from validation dataset are used to assess the model performance. Usually when the area under ROC curve (AUC) is greater than 0.8, the model can be considered “good”. Our model has an AUC of 0.82. Figure 3.5 measures the modeling performance in another way, cumulative gains chart. Based on Figure 7, we can see that top 5% of the hand-raisers capture about 47% of the escalations; top 15% of the hand-raisers capture about 77% of the escalations.

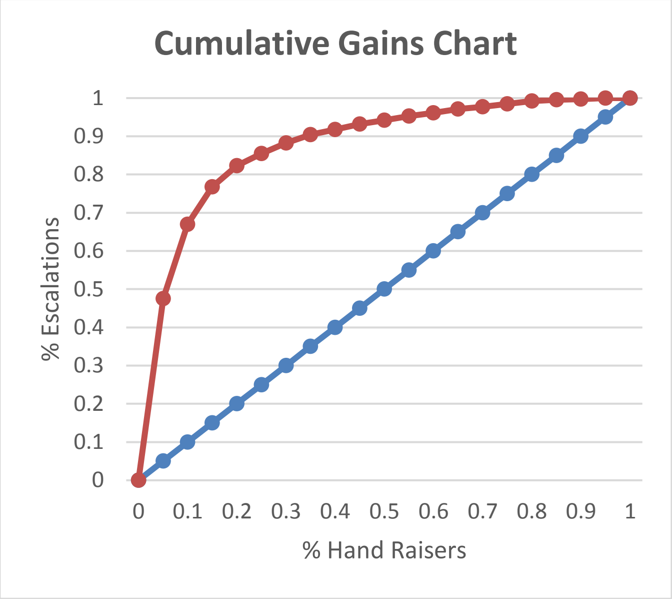


Figure 7: Gain Chart - % “Hand Raisers” versus % Escalations

After building predictive model, the results are combined in order to rank the risk score or probability of escalation for the whole population. The purpose of this study is to target which groups of customers have high probability of escalation. In order to make business implementation meaningful, we group the customers into 5 risk levels which are high, med-high, medium, med-low and low. Table 1 summarizes the results and Figure 8 visualizes out findings with bar chart.

Table 1: Customer Risk Level and Corresponding POE

Risk Level	**%of Hand Raisers**	**% of Escalations**	POE Range
High	5	47	>= 0.17
Medium High	10	30	(0.053, 0.17)
Medium	25	15	(0.0169, 0.053)
Medium Low	30	6	(0.0063,0.0169)
Low	30	2	<=0.0063

The beauty of the table and chart is that it aligns with 20/80 rule, which is 20% of total population makes 80% contributions. The results can be directly applied into business process. If a customer calls in to complain about the vehicle, their probability of escalation can be immediately calculated. If he/she falls in high or med-high group, a red flag should be put right away. Customer retention management teams can ask dealer to take good care of him/her, and offer some goodwill. Thus, the company can benefit a great deal about customer retention management from the predictive model.

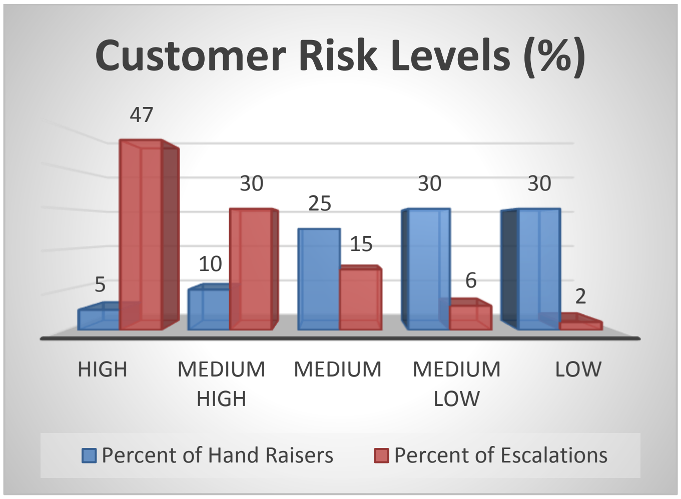


Figure 8: Segment Customers by Risk Levels

Significant predictors also deserve to dig deeper. Table 2 below summarizes the significant variables by their polarities. Here number of "+" and "-" denotes the importance of the positive and negative predictor variables, respectively. The rationale explanation of each predictor is also discussed here, and all of them make business sense. Based on these predictors, business team can develop new policies to improve customer retention management.

Table 2: Predictors Summary and Rationale Explanation

	Predictors	Rationale Explanation
Positive (Variable increases POE)	Number of High Severity Claims (+++)	Severe claims make escalation more likely
	Total Vehicles Purchased (++)	More vehicles make escalation more likely; Sense of entitlement for being a loyal customer
	Early Failure Period (++)	Problems more likely to occur early on
	Number of Cases (++)	More cases make escalation more likely
	Pacific (++) Midwest (+)	Californian Laws
	First Year of Model (+)	New models are less reliable
	Maximum Days Down (+)	Longer waits anger customers
Negative (Variable increases POE)	Longevity (--)	More attached to brand making escalations less likely
	Percent of Total Vehicles with Additional Warranty (-)	Problems under warranty can be earlier and easier to address
	Percent of New Vehicles in Purchase History (-)	Certified used vehicles more likely to have problems
	Dealer Rating Score (-)	Better dealers take better care of customers

IV. Conclusion

This study presents a practical process to build a reliable data structure for big data analytics and predictive modeling. Although automotive industry data is used to illustrate the usefulness of this approach, it can be applied into any industry with large amounts of customer data. Comprehensive data about customers is collected from four subject categories, but the structure that has been developed is scalable and extensible as other useful business facts are discovered. Using logistic regression technique, significant predictors that drive the POE are identified and relative importance is quantified. The modeling results show that the data mining techniques can help companies allocate limited resources based on facts, rather than intuition, to serve better customer retention management.

References

[1] Liao, P. (n.d.). Hadoop Family and Ecosystem. Retrieved January 14, 2016, from http://www.slideshare.net/tcloudcomputingtw/hadoop-family-and-ecosystem-15693655

[2] Pogol, Gina. Tips for Cost-Effective Customer Retention Management, www.crm2day.com/library/docs/50577-0.pdf, 2007