

**What Drove RV Insurance**

**Purchasing Behaviors in India?**

**STAT701 Modern Data Mining - Final Project**

**Anirudh Bajaj, Raman Chadha, Matthew LeBaron, Esther Shin**

**Executive Summary**

The RV industry has gained healthy momentum in the past few years, fueled by the Millennial generation’s enthusiasm for travel. A complementary industry that in turn experienced significant growth is the RV insurance industry, for which statistical predictive analytics has proven valuable over the course of years. The dataset that our team used to demonstrate the usefulness of such analysis was gathered in India a few years ago. The goal of this analysis is to predict what kind of individual is most likely to purchase insurance policies, to ultimately help with insurance firms’ decision making and risk controlling practices in marketing, pricing, and underwriting.

We conducted our analysis by first using Exploratory Data Analysis to understand and groom the data. We interpreted the various variables and plotted them to check if there was sufficient variability in the data and if the variable was connected to the output we were trying to predict.

Next, we experimented with multiple different model types including glm, backward selection, lasso and elastic net, and random forest to see which model could give us the best predictions going forward. We found that the logistic model with backward selection produced comparable AUC and MCE results while limiting the number of predictive variables (and thus reducing the risk of overfitting).

**The most important variables in predicting RV insurance purchases were other types of insurance purchases such as car and boat insurance, as well as education level and marriage status.**

Although these models are fairly predictive, we would exercise caution when applying the results to other geographies or markets. Additionally, data limitations such as sample size, importance of insurance branding, insurance price points, and sampling time frame could impact attempts to leverage these models in other settings.

**Project Background and Goal**

It is well-known that Millennials love to travel, so much so that this generation is expected to have more impact on the travel and tourism industry than any other preceding generations had. One of the segments that the millennials have recently reinvigorated is the RVs industry. In fact, CNBC has reported that ‘the RV space is on fire’ as the millennials are expected to push sales to record highs. Major industry players such as Winnebago and Thor have reported record growth, and the trend is expected to continue. Our team, also consisting of avid traveler millennials, found this trend highly intriguing and decided to apply classroom concepts to a related real world problem.

The dataset that we selected for the final project focuses on caravan (RV) insurance policies, taking place in India a few years ago. Given that this dataset represents a practice in a different time and different place, our team was intrigued by how the results would differ from what we now see in the current-day United States.

**The goal of our analysis is to predict who would purchase mobile home (caravan) insurance in India, based on demographic and behavioral factors of the potential customers.** This analysis when applied in real-world settings would be highly beneficial to any RV insurers in controlling risks especially in the marketing, pricing, and underwriting processes.

**Data Summary and Exploratory Analysis**

Data Introduction:

Initial data that we obtained from Kaggle contained 9822 observations across 87 variables. The row data contains unique potential customer data who may or may have not purchased the caravan (RV) insurance. The columns provide robust information about unique potential customers, including:

* Demographic factors: Containing information across age, religion, marital status, presence of children in the household, social status, income, home ownership, car ownership, etc.
* Insurance purchasing factors: Containing information on the number of insurance policies purchased across various non-RV coverage areas (e.g. car, delivery van, life insurance, disability insurance, social security insurance, etc.) and level of contributions to such insurance policies

***Please see Appendix for a detailed explanation of each variable in the dataset***

Data Preparation and Exploration

We started with missing values examination to see if any columns or rows had NA values, and deleted rows with missing values. In addition, we excluded unique ID variable without predictive power from the analysis to reduce the complexity of the model.

We then examined the response variable (number of caravan/RV insurance), and learned that about 20% of the population within the dataset purchases a mobile policy, while the remaining two thirds did not.

In order to determine which input variables should be included or omitted to predict how many insurance purchases are bought, we performed analysis on skewness/variability and correlation across demographic and insurance purchasing behavior variables (and corrected/converted each variable to the most appropriate forms as needed).

First off, Customer Main Types and Customer Subtypes were converted into factors, and were plotted against the Number of Caravan Policies (policy.output) to check for any significant variation. There was reasonable variation across both variables hence they were left untouched.

Same analysis was performed for Size of Household variable, which showed reasonable variation and was left untouched as well.

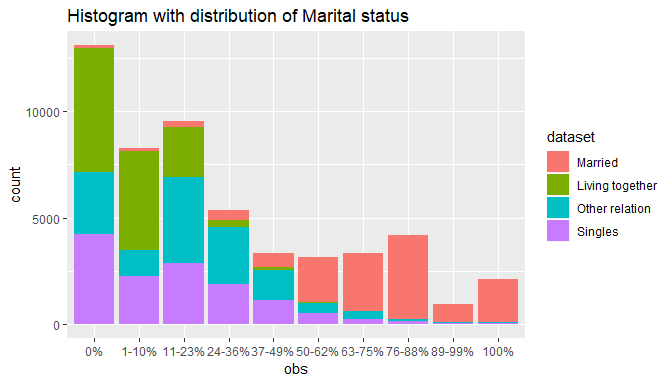
Next variable that was tested with variance was Number of Houses. We saw that the data was highly skewed, and could benefit from a log transformation, but since the number is so low (10), we made a decision to ignore the transformation to reduce complexity.

We also took a close look at the Age variable. Looking at the structure of the data, we realized that these are factors as per the table, representing different brackets of age groups. We decided that it would be helpful for the interpretation to rename the variable's levels, because factor levels (1-6) might be prone to misinterpretation. Variability was tested, and the Age data was approximately normal, passing the test. Same analysis was conducted across Income, Purchasing Power, Religion, Marital Status, Education, Social Status, Social Class variables, all of which showed reasonable variations.

In addition, we conducted high-level correlation analysis to test if any input variables were significantly correlated based on a selection of input variables. We concluded that key variables had mostly weak correlation to each other, although some were more correlated to each other than another especially between Income and other demographic variables that indicated financial status.

For variables that were related to each other, we plotted them together on a graph to check which variables were connected to the output variable, accounted for the variability, and to draw high-level insights from the data which could be used for model building.

For example, in the case of the 4 variables related to Marital status, we found the below graph:



We can clearly see that all 4 variables are distributed across the factors, and married people tend to live across each other in similar neighbourhoods.

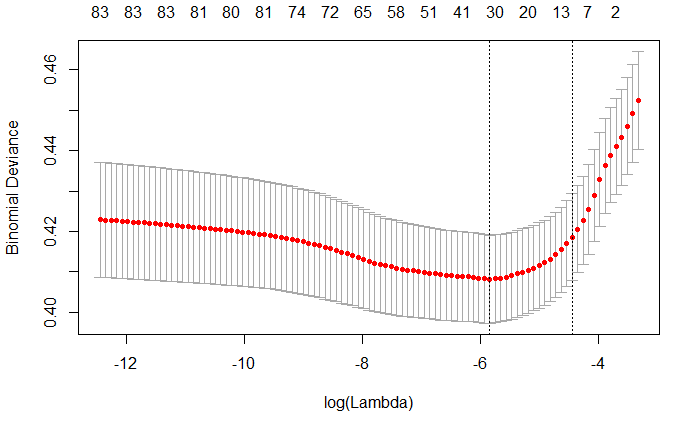
At the end of the EDA section, we eliminated only the identification variable from the analysis, and other variables were considered to be considered for the first model building set.

**Detailed Analysis and Model Building**

*LASSO and GLM*

After Exploratory Data Analysis, we begin the model-building process. First, we prepare a design matrix and response column. Next, we begin variable selection. One option is to use all variables to build the model; however, to avoid overfitting the data, we choose to perform lasso techniques and elastic net.

We start by setting the seed so that our analysis can be reproduced in the future (as opposed to outputting different variables each time, which would require a new model to be built). We then use ‘cv.glmnet’ with 10 fold cross-validation from which we can choose a lambda.min value, visualized in the following chart:



After applying the lambda.min filter, we find 32 variables (out of the 86 total variables) which are most predictive, without introducing unnecessary variables. These 32 variables are subsequently used to predict caravan insurance purchases using a ‘glm’ model, which produces the following results:

Variables with high significance (<0.001 p-value):

* Car insurance purchases
* Fire insurance purchases
* Number of boat insurance policies

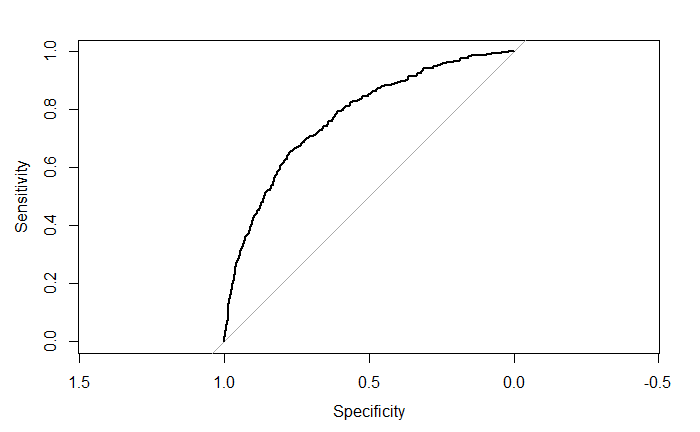
Variables with medium significance (0.01 to 0.05 p-value):

* Person does not work as a farmer
* Disability insurance purchases

Variables with low significance (0.05 to 0.1 p-value):

* Person works in middle management
* Person does not have income above 123,000
* Private third party insurance purchases

The Area Under the Curve (AUC) produced by this model is 0.77 and is modeled in the chart below:



Additionally, we will incorporate a **loss function adjustment of 0.2** since we are much more comfortable marketing to those who are less likely to purchase than missing people who may have bought insurance. In other words, our overage cost is much less than our underage cost. So we make the adjustment by setting any records with a score above 0.2 to be “1” and any records with a score below 0.2 to be “0”. This loss function broadens our marketing base while still focusing on prospects who are most likely to respond.



a0,1/a1,0 = ¼ = 0.25

Hence, probability threshold = 0.25/(1+0.25) = ⅕ = 0.20.

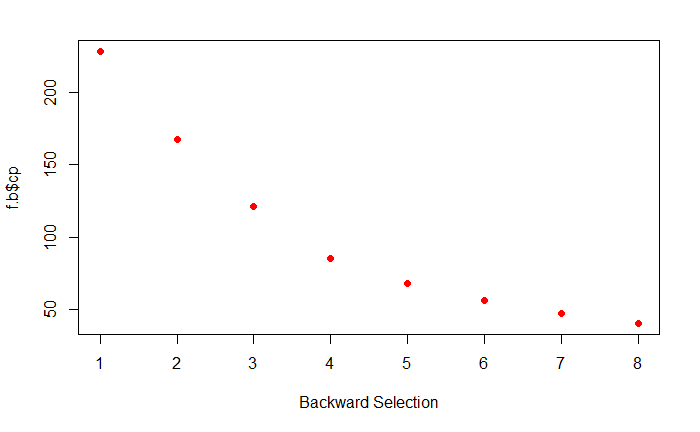
Finally, we find that our model has a weighted misclassification error (MCE) of 0.27.

*Backward Selection*

In addition to lasso selection, we perform backward selection to compare models and see if we are getting similar outputs. Using all variables as predictors, we find this version of the eight most important variables in this order:

1. Did this person purchase car insurance?
2. Did this person purchase boat insurance?
3. Does this person have a lower level education?
4. Did this person purchase fire insurance?
5. Is this person married?
6. Did this person purchase third party agriculture insurance?
7. How many social security insurance policies does this person have?
8. Is this person a farmer?

These variables are very similar to those obtained through the lasso / glmnet method, which is what we would expect. Also, the cp values drop off well by the eighth variable, as shown in this chart:



Thus, we find that backward selection is another valuable approach to producing a basic model with the most important variables (but not so many variables that we overfit the data).

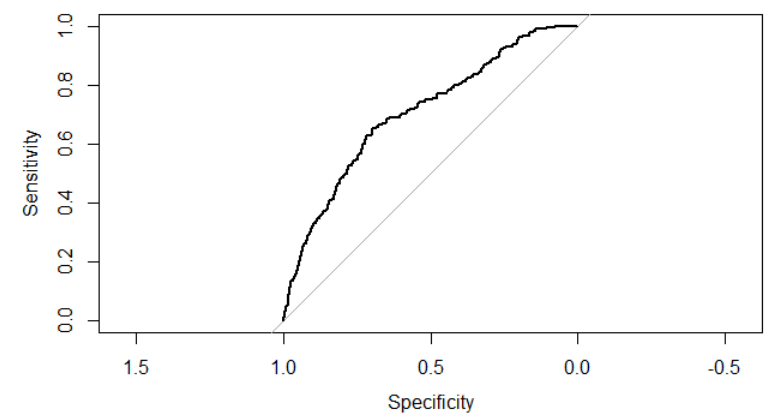
*Random Forest*

We use the Trees approach to validate the output and potentially look for a better performing model. Since a single classification tree does not give appropriate results, we move to Random Forest technique.

We first use the randomForest package with CARAVAN as the response variable and all other variables as predictors. We get an AUC of 0.6957 and Testing error of 0.062. The output chart suggests a minor improvement by using 500 trees, which is what we use hereafter. Since randomForest package functions by using the “Majority Vote” methodology, it does not allow us to customize the loss function as per the requirements.

We then use “ranger” package to overcome the above challenge, and use the splitrule as “gini”. We use the default mtry = 9 value as suggested by randomForest. We first run the ranger function on the entire data-set and use Out Of Bag (OOB) error to estimate the prediction error. We get an OOB error of 0.065, which is very close to the testing error above. We also validate this against our testing data, where we get the Testing error as 0.062 (as expected, this is same as that given by randomForest package). We get a slightly improved AUC of 0.709 in this case. Finally, we apply the revised Loss Function to check for the mis-classification error. We get a Testing error of 0.100, a slight increase over the default majority vote Testing error.

Final Random Forest ROC with an AUC of 0.709:



*Model Comparison Summary:*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Model** | **# of Variables** | **AUC** | **Misclassification Rate\*** |
| **1** | **Full Logistic** | **85** | **0.79** | **0.26** |
| **2** | **Logistic with Backward Selection** | **8** | **0.76** | **0.27** |
| **3** | **LASSO / Elastic Net** | **32** | **0.77** | **0.27** |
| **4** | **Random Forest** | **85** | **0.71** | **0.10** |

While recommending the final model, besides AUC and Testing Error (Misclassification rate), we also consider interpretability and model parsimony as important factors. Hence, we recommend logistic with backward selection as the final model as it is good fit overall and uses the least number of variables to generate a prediction approximately as powerful as the other models.

**Limitations**

1. The data was gathered in the previous decade and the use cases and prevalence of RVs might have evolved over the years. The insurance industry itself might have undergone some changes that are not captured by this analysis.
2. Our level of confidence in this analysis replicating in the India market is fairly high. However, there could be other important factors as well for different countries.
3. Interaction effects are not considered to keep the model simplistic. The number of variables at the outset itself make the analysis difficult to consume. It is possible that certain predictors influence certain category of categorical predictors more or less than the others. Such effects have not been explored in this study in detail.
4. The data set provides limited insight into how the data was collected by Sentient Machine Research in the first place. Prima facie, it is difficult to ascertain how certain categorical predictors were made mutually exclusive. For example, there might be some overlaps among the “Customer Subtypes”, such as Stable Family and Affluent Young Family etc.
5. Insurance purchase behavior is likely to also depend on certain insurance company parameters, such as price points, benefits, premium, coverage, competition, spend on advertisements, modes of advertisements etc. These have not been incorporated in this analysis.

**Appendix**

**Data Explanation**

The data file contains the following fields:

* ORIGIN: *train* or *test*, as described above
* MOSTYPE: Customer Subtype; see L0
* MAANTHUI: Number of houses 1 - 10
* MGEMOMV: Avg size household 1 - 6
* MGEMLEEF: Avg age; see L1
* MOSHOOFD: Customer main type; see L2

\*\* Percentages in each group, per postal code (see L3)\*\*:

* MGODRK: Roman catholic
* MGODPR: Protestant ...
* MGODOV: Other religion
* MGODGE: No religion
* MRELGE: Married
* MRELSA: Living together
* MRELOV: Other relation
* MFALLEEN: Singles
* MFGEKIND: Household without children
* MFWEKIND: Household with children
* MOPLHOOG: High level education
* MOPLMIDD: Medium level education
* MOPLLAAG: Lower level education
* MBERHOOG: High status
* MBERZELF: Entrepreneur
* MBERBOER: Farmer
* MBERMIDD: Middle management
* MBERARBG: Skilled labourers
* MBERARBO: Unskilled labourers
* MSKA: Social class A
* MSKB1: Social class B1
* MSKB2: Social class B2
* MSKC: Social class C
* MSKD: Social class D
* MHHUUR: Rented house
* MHKOOP: Home owners
* MAUT1: 1 car
* MAUT2: 2 cars
* MAUT0: No car
* MZFONDS: National Health Service
* MZPART: Private health insurance
* MINKM30: Income < 30.000
* MINK3045: Income 30-45.000
* MINK4575: Income 45-75.000
* MINK7512: Income 75-122.000
* MINK123M: Income >123.000
* MINKGEM: Average income
* MKOOPKLA: Purchasing power class

\*\* Total number of variable in postal code (see L4)\*\*:

* PWAPART: Contribution private third party insurance
* PWABEDR: Contribution third party insurance (firms) ...
* PWALAND: Contribution third party insurane (agriculture)
* PPERSAUT: Contribution car policies
* PBESAUT: Contribution delivery van policies
* PMOTSCO: Contribution motorcycle/scooter policies
* PVRAAUT: Contribution lorry policies
* PAANHANG: Contribution trailer policies
* PTRACTOR: Contribution tractor policies
* PWERKT: Contribution agricultural machines policies
* PBROM: Contribution moped policies
* PLEVEN: Contribution life insurances
* PPERSONG: Contribution private accident insurance policies
* PGEZONG: Contribution family accidents insurance policies
* PWAOREG: Contribution disability insurance policies
* PBRAND: Contribution fire policies
* PZEILPL: Contribution surfboard policies
* PPLEZIER: Contribution boat policies
* PFIETS: Contribution bicycle policies
* PINBOED: Contribution property insurance policies
* PBYSTAND: Contribution social security insurance policies
* AWAPART: Number of private third party insurance 1 - 12
* AWABEDR: Number of third party insurance (firms) ...
* AWALAND: Number of third party insurance (agriculture)
* APERSAUT: Number of car policies
* ABESAUT: Number of delivery van policies
* AMOTSCO: Number of motorcycle/scooter policies
* AVRAAUT: Number of lorry policies
* AAANHANG: Number of trailer policies
* ATRACTOR: Number of tractor policies
* AWERKT: Number of agricultural machines policies
* ABROM: Number of moped policies
* ALEVEN: Number of life insurances
* APERSONG: Number of private accident insurance policies
* AGEZONG: Number of family accidents insurance policies
* AWAOREG: Number of disability insurance policies
* ABRAND: Number of fire policies
* AZEILPL: Number of surfboard policies
* APLEZIER: Number of boat policies
* AFIETS: Number of bicycle policies
* AINBOED: Number of property insurance policies
* ABYSTAND: Number of social security insurance policies
* CARAVAN: Number of mobile home policies 0 - 1

### Keys (L1 - L4)

L0: Customer subtype

* *1*: High Income, expensive child
* *2*: Very Important Provincials
* *3*: High status seniors
* *4*: Affluent senior apartments
* *5*: Mixed seniors
* *6*: Career and childcare
* *7*: Dinki's (double income no kids)
* *8*: Middle class families
* *9*: Modern, complete families
* *10*: Stable family
* *11*: Family starters
* *12*: Affluent young families
* *13*: Young all american family
* *14*: Junior cosmopolitan
* *15*: Senior cosmopolitans
* *16*: Students in apartments
* *17*: Fresh masters in the city
* *18*: Single youth
* *19*: Suburban youth
* *20*: Etnically diverse
* *21*: Young urban have-nots
* *22*: Mixed apartment dwellers
* *23*: Young and rising
* *24*: Young, low educated
* *25*: Young seniors in the city
* *26*: Own home elderly
* *27*: Seniors in apartments
* *28*: Residential elderly
* *29*: Porchless seniors: no front yard
* *30*: Religious elderly singles
* *31*: Low income catholics
* *32*: Mixed seniors
* *33*: Lower class large families
* *34*: Large family, employed child
* *35*: Village families
* *36*: Couples with teens 'Married with children'
* *37*: Mixed small town dwellers
* *38*: Traditional families
* *39*: Large religous families
* *40*: Large family farms
* *41*: Mixed rurals

L1: average age keys:

*1*: 20-30 years *2*: 30-40 years *3*: 40-50 years *4*: 50-60 years *5*: 60-70 years *6*: 70-80 years

L2: customer main type keys:

* *1*: Successful hedonists
* *2*: Driven Growers
* *3*: Average Family
* *4*: Career Loners
* *5*: Living well
* *6*: Cruising Seniors
* *7*: Retired and Religeous
* *8*: Family with grown ups
* *9*: Conservative families
* *10*: Farmers

L3: percentage keys:

* *0*: 0%
* *1*: 1 - 10%
* *2*: 11 - 23%
* *3*: 24 - 36%
* *4*: 37 - 49%
* *5*: 50 - 62%
* *6*: 63 - 75%
* *7*: 76 - 88%
* *8*: 89 - 99%
* *9*: 100%

L4: total number keys:

* *0*: 0
* *1*: 1 - 49
* *2*: 50 - 99
* *3*: 100 - 199
* *4*: 200 - 499
* *5*: 500 - 999
* *6*: 1000 - 4999
* *7*: 5000 - 9999
* *8*: 10,000 - 19,999
* *9*: >= 20,000

**R code section is included as a separate (.Rmd) attachment**