Competition Spring 2016

Predict a policy type and price for a customer based on browsing and transaction history

INFO 7309 Machine Learning for Business Intelligence

Team: **Data Wizards**

Leela Gangadhar Vallabhaneni

Nuhiya Rafeeq

Sumit Deshmukh

Department of Information Systems,

Northeastern University

Date: 04/27/2016

**Project Understanding:**

* + Dataset related to Car insurance
  + Our Target:   
    1. Predicting policy which is a classification problem.   
    2. Predicting cost which is a Regression problem
  + Cost of a product is a function of both the product options and customer characteristics.
  + A customer may purchase a product that was not viewed.

**Approaches**

* Approach 1: Non-hypothesis driven data analysis (i.e. Boiling the ocean)
  + A lot of data exists and without properly understanding the data, trying to run some analysis.
  + Time wasted analyzing and understanding every available variable which can be reduced through proper data understanding.
  + We need to understand what we are looking for i.e, meaningful attributes that make sense to the target.
  + Once there is enough insight into the problem, explore every possible variable and relationship in hope to use all that shows promise.
* Approach 2: Hypothesis driven analysis
  + List down comprehensive analysis
  + See if they are available.

**Hypothesis Generation**

* What?
  + Hypothesis is a possible view about the problem. Can be anything that makes sense.
  + May be true or may not be true. We will get answers after data understanding whether it needs more attention.
* Why?
  + Building features which are not biased by the data available in the data-set. Don’t just focus on the data that’s available directly. There is a good chance to produce new features from existing features if there is a good hypothesis.

**Some of the Hypothesis made:**

* Does cost depend on policy?
* Does the financial status depend on the policy purchased?
* Do customers with a risky profile tend to buy a certain kind of policy?
* Does the state matter?

This is **Business understanding** where the problem and resources needed for possible solutions are discussed.

**Data Understanding**

* Different Types of Attributes

|  |  |  |
| --- | --- | --- |
| **Categorical** | **Ordinal** | **Numerical** |
| State | Policy | Group Size |
| Location | Risk Factor | Homeowner |
| Married Couple | Age\_Oldest | Car Age |
| Car Value | Age\_Youngest | Duration\_Previous |
| Day | C\_Previous | Cost |
|  | Time |  |

* Does State plays an important role?
* How to make use of session data?
* Did the Customer change C\_previous in later sessions?
* When did he change the C\_previous, at which shopping point?
* How much did cost changed due to the change in C\_previous
* Did he view the policy that he purchased?
* If the browsing sessions could be combined into one session per customer, would this help generate a new attribute?

**Data Preparation**

* Different Types of attributes available in the dataset which need to be converted to numeric types at the end.
* Duplicates where row data is same in session data. Session data where rows except time is similar for every customer.
* Missing Values, NA values, Outliers.  
  Since there is about 30% of rows with missing or NA values, we can’t just discard them. So we are using MICE to impute missing values.
* Different ranges of attributes like time are converted to categorical values using specific range.
* Unbalanced classes exist in our dataset. Policy 1 and 3 itself are taking 70% of the data. So we are using oversampling using SMOTE to create synthetic data.

**What is the difference between Categorical, Ordinal and Interval variables?**

* **Categorical**
  + (Also called Nominal Variable)
  + No intrinsic ordering to the categories
  + Ex: Gender: male and female; no order
* **Ordinal**
  + Clear ordering of the variables
  + Economic status = (low, medium and high)
  + Type of categorical, but they can be is ordered
* **Interval**
  + The intervals between the values of the interval variable are equally spaced
  + Income of three people is 1000, 2000, 3000. Size of the interval is same.

**Why does it matter whether a variable is Categorical, Ordinal or Interval?**

* Statistical computations and analyses assume that the variables have a specific levels of measurement (numeric).
* An average of a categorical variable does not make much sense (Ex: average of color Blue).
* Average of an ordinal variable such as economic status would also obtain a nonsensical result.

**What to do with Categorical Variables?**

* Since they do not have any particular order, they cannot be treated as numeric variables.
  + Ex: Day, State, Location, Car\_Value, etc
  + Solution:
    - One hot encoding/ 1-of-n encoding.  
      Convert single column into the categorical number of columns with binary classification.

**Duplicate Records**

* Train.csv and test\_session\_history.csv have session history of customer.
* For a customer multiple session records with same details are considered as duplicates. These have been deleted.

**Missing Values**

* Risk Factor: MICE
  + Multivariate Imputation by Chained Equations is used to impute missing values with multiple imputations.
  + Imputes missing values with plausible data values.
  + Why? –Since Risk Factor depends on customer characteristics it is important to using all the data rather than replacing missing with mean or median.
* C\_Pervious and Duration\_Previous
  + Median values.
  + Why? – Since there is no dependence on other features in the dataset, they can be replaced using standard statistical results.
* Car Value in train.csv
  + Median is used to replace 1000 missing values.

**Outliers**

* Measurements significantly distant from other observations.
* Clip Values task in Microsoft ML Studio.

**Normalization**

* What?
  + Adjusting values measured on different scales to a common scale, in this case a Normal Distribution.
* When?
  + Various attributes having various ranges
* Why?
  + This can guarantee stable convergence of weight and biases.
* **How?**
  + **standard score (**Z-score)
  + (a) allows to calculate the probability of a **score** occurring within our normal distribution
  + (b) enables comparison of two **scores** that are from different normal distributions

**Cross Validation**

* What?
  + Also called Rotation Estimation.
  + A method of assessing the accuracy and validity of a statistical model.
  + Achieve more generalized relationships.
  + Did not yield significantly better results.

Using 10 fold Cross validation.

**Under Sampling**

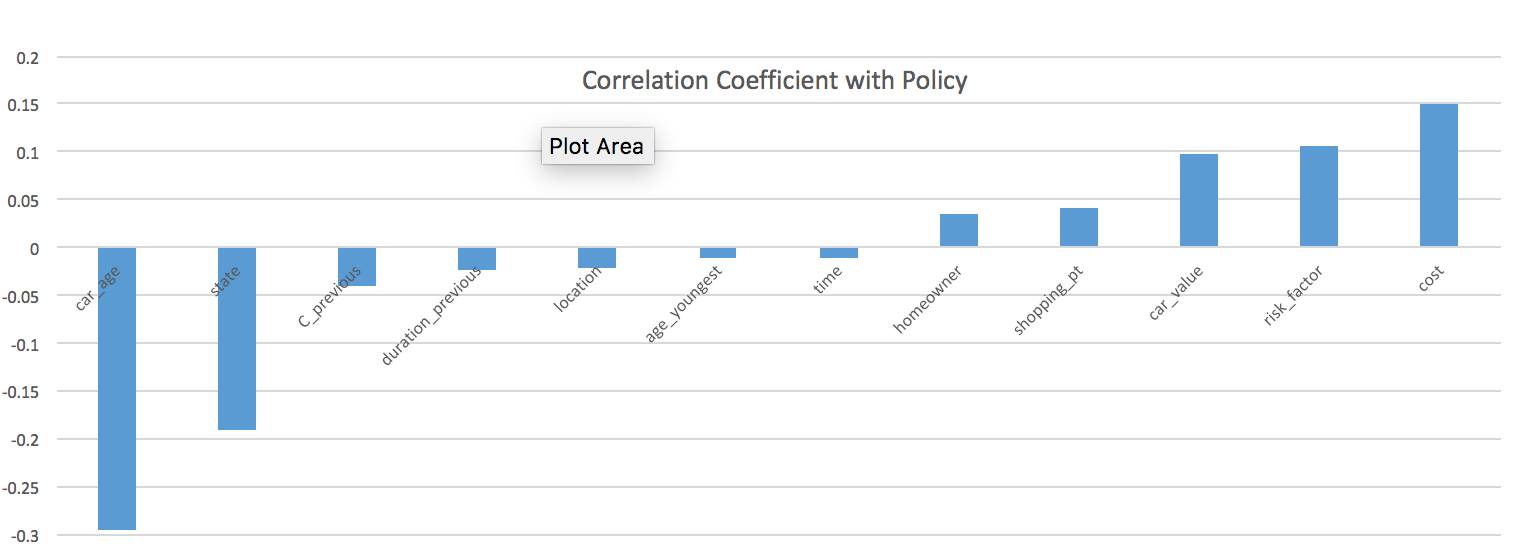
* Use fewer samples of highly represented classes. Reduced the number of rows Policy 1, 2 and 3 to ranges of policy 4.
* Decision Forest Algorithm is used for classification.
* It produced and accuracy of 75.74%.
* Result: Decreased Accuracy.

**Over sampling using SMOTE**

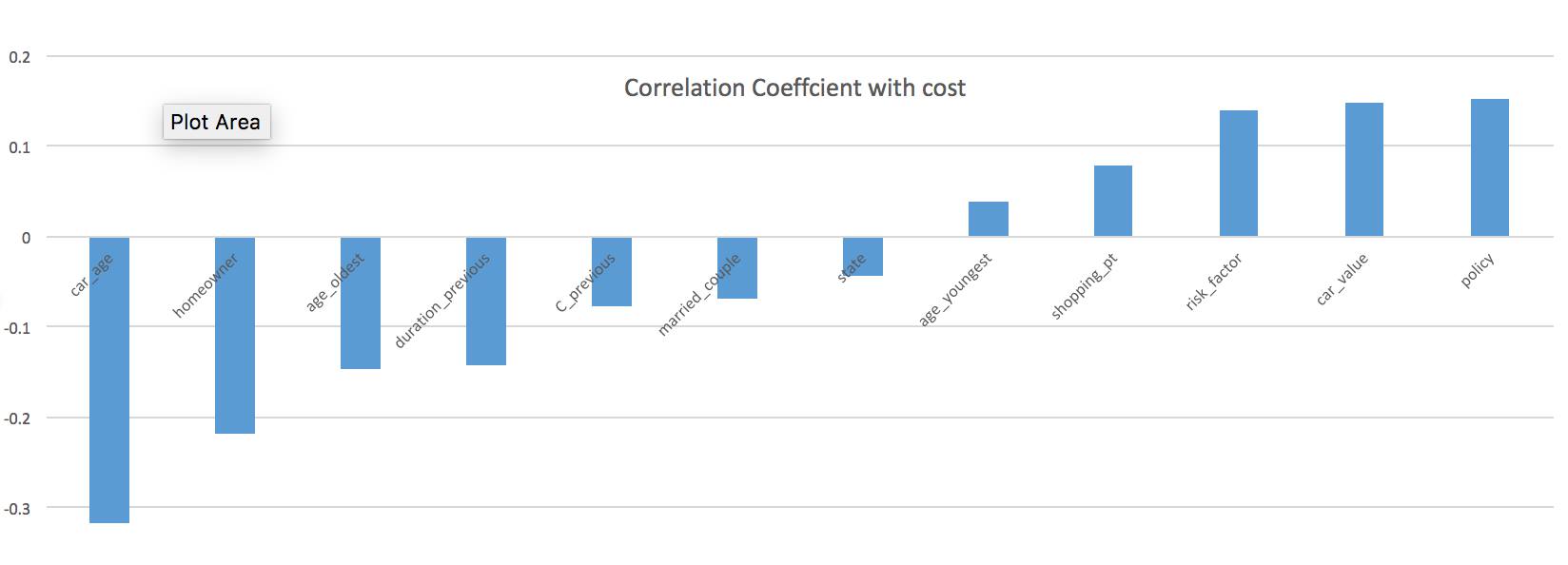
* Synthetic Minority Over-sampling Technique.
* Used for over sampling data.
* Synthetic data is introduced for low represented classes.
* Azure ML SMOTE is used.
* Policy No-2 rows have been increased.
* Policy No-4, still poor number of rows.

**Feature selection**

Correlation coefficient of every feature Vs Policy



Correlation coefficient of every feature Vs Cost



Selecting the top features for Policy and Cost separately is important because both have different correlated features.

**Using low correlated features with the help of PCA**

* Feature reduction
* Variables not closely correlated.
* PCA on:
  + Married couple
  + age\_youngest
  + age\_oldest
  + shopping\_point
  + group\_size
  + duration\_previous.

