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: 04/13/2016

Approach

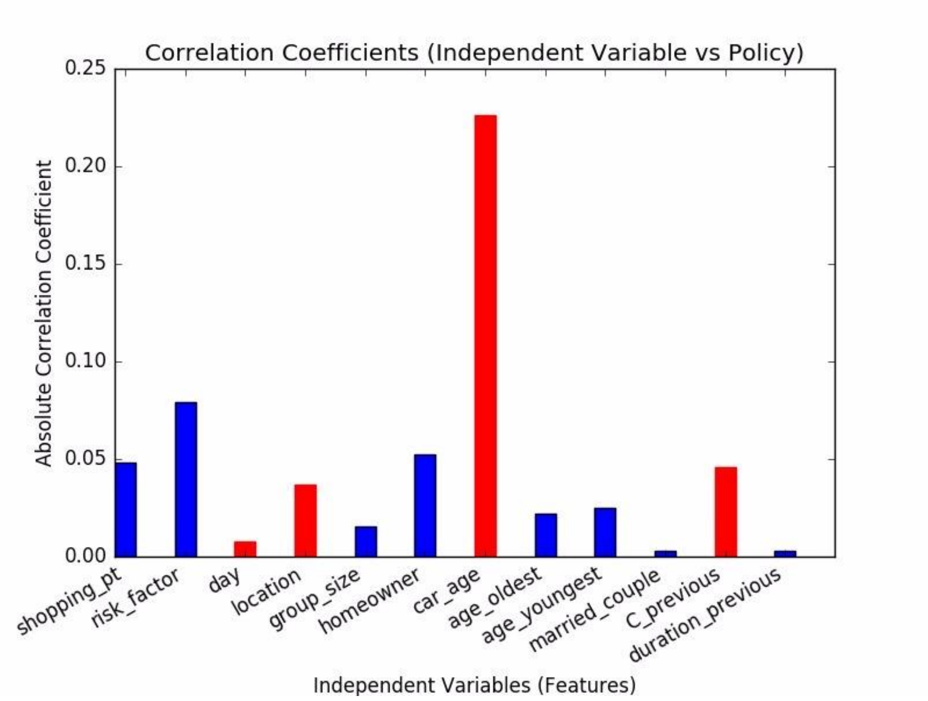
The problem is divided into 2 steps. Step 1 is to predict the policy number by using classification techniques (Decision tress, SVM and KNN) using the train_short.csv. Step 2 is to predict the policy price using Linear Regression, SVR and Random Forest methods.

Statistical Findings

The correlation coefficient was calculated for the purpose of feature selection. Top 6 features with the maximum Correlation Coefficient- car age, risk_factor, shopping_pt, home owner and C_previous were selected to train the model.

Pre-processing of Data

The Data was preprocessed by eliminating the NA values(~30% of data set). No missing values and duplicate records were found. To convert categorical attributes into numerical attributes, we are going to apply one-hot encoding method.



Outcomes for Step1: Predicting the Policy number

Using Random Selection Train-Data and Test-Data have been split as 80:20.

1. Model-1, Decision Trees: Decision Trees performed very low.

Train Data	Train Data
• Accuracy 0.477	• Accuracy 0.477
• Precision 0.34	• Precision 0.34
• Recall 0.48	• Recall 0.48

2. Model-2, SVM: SVM has high accuracy on the Train Data and low accuracy on test data

Train Data	Test Data
• Accuracy 0.955	• Accuracy 0.57
• Precision 0.97	• Precision 0.58
• Recall 0.96	• Recall 0.55

3. Model-3, KNN: KNN turned out to be a better model.

Train Data	Test Data
• Accuracy 0.70	• Accuracy 0.56
• Precision 0.7	• Precision 0.55
• Recall 0.71	• Recall 0.57

Observations

- The data was found to be imbalanced with very fewer entries for policy numbers 2 (20%) and 4 (5%). The model was biased to the policy numbers 3 and 1.
- Smaller class were being treated as an outlier and the model learns the decision boundary for only the larger class.
- Over-sampling or under-sampling techniques to be applied.

Up Next,

- Techniques for predicting policy price are being explored.
- Linear regression, SVR and Random Forest methods will be used to predict the Policy Price.