Edmunds.com Traffic Data Exploration, Group name: NeuralSpyder

Data Review: The data provided by Edmunds.com is composed with the online-generated traffic data and the final purchasing records corresponding to the visitors (not unique), potential purchasers (the visitors who make leads) and the final purchasers who had the online activity in the real auto market. Our model is to define the potential elements on the website that can drive the visitors to become potential purchasers. This intuition can help the website to improve the user experience and accommodate the needs of customers and dealers. Moreover, the dealers can also make adjustments to find potential customers.

Data Cleaning: The data should be prepared to accommodate the further data analysis and our goals. The data set needs to be merged on the unique customer key. The model clarifies the customers into two types: the online visitors only (0) and the online visitors who become purchasers later in auto market (1). We prepared each of the customers so that they have the same variables. Basically, we can analyse the offsets among the variables between the two categories. The missing data are random imputed by their own data in each column. The categorical data and the numerical data should be also properly handled.

Unsupervised Learning: Correlations of the features (variables): Consider each feature(variable) as a vector X_i, we have the covariance: $cov(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))]$, so that we can plot a covariance map to see the covariance between each and every pair of the features (variables). That implies whether these pairs of features are correlated with each other. The covariance map of the visitor table is as follows: Filtered out the covariance above 0.9, we have the covariance map:
That implies the following pairs of variables are coupled:

Group 1: (The total count of the pages viewed in each categories)
(consideration_count, configuration_count)
(configuration_count, nci_count)
(configuration_count, new_leads_count)
(nci_count, new_leads_count)

In the above way, we have revealed the intrinsic structure of the variables - that suggest that which variables are coupled together and thus is an indicator of how users use the website. Interestingly, we found that people who tend to take a glance at the configuration category of web pages, they tend to join the leads. However, the consideration - leads relationship is not comparable as the former one. So, if we can turn the customer to engage in the configurations of their wanted car, we may get a step further to make them buy a car after all.

Supervised Learning: Boosting Prediction Model: Build a prediction model so that the website can define the possible purchasers according to the traffic generated by the users: the model can give the most influential factors and evaluate the prediction result by regressing on validation dataset. The response variable is a binary variable: 1 represents the visitors who become purchase and 0 represents that pure visitor without purchasing at the end. There are 400 predictors originally, but we select the most significant 9 predictors by forward data selection by the criteria of cross validation in fitting the logistic model. Boosting is the final chosen prediction model: the bagging and random forests does not have a significant binary-data prediction. The top three variables selected by the boosting are new_dwell_time, pc_new_mydp and consideration_count. Then the model can also generate the partial dependence plots for the top three important variables. These three plots illustrate the marginal effect of the selected variables on the response after integrating out the other variables: the visitors are more likely to becomes a purchaser if new_dwell_time is increasing with the other variables be constant. The error rate by fitting on the validation dataset is 0.1717305 according to the confusion table. This is a relatively good prediction for our model.

Recommendation: The website needs improve the customer experience so that it can make the customer stay on the websites and view as more pages as possible. Particularly, the website has good margin of growth for the user experience of the users who use the PC to login the website.

Further Improvement:Here is a list potential improvements the model can make in the future: 1. perform more sophisticated data transformation and missing data imputation to avoid biases and explore the underlying pattern of the data more accurately. 2. Carry on Propensity Score Matching to avoid the imbalance variables so that can make causality analysis to define the most influential predictors. 3. Categorize the customer into more sophisticated clusters, for example, the visitors can be broken down into three categories: the online visitors that do not have willingness to buy, the visitors that have made leads and the visitors make leads and become purchasers. At least but not limited to this, the the further model can also be clustered in terms of a group of variables. 4. Perform Bayesian Hierarchical Clustering model to define the customer groups. Carry one Bayesian Discrete Choice model to analyze the distributions of parameters across different clusters, so that the website target on the parts that are most likely to drive the visitors to become potential buyers. 5. Work with DBA to adjust the definition for each of the variable generated by the website, for example, the estimation of the daily unique visitors. Obtain related influential variables, such as the website taking the customer into the Edmunds.