Geographic Heat Map For Buying Price Based on Random Forest Predictions

The Sunsets: Ouwen Huang - Joy Patel - Yu Zhou Lee - Josh Miller

Our goal for this project was to visualize and predict by state the final price of a car that a user buys from Edmunds. This data set has a lot of information about individual users, and aggregate statistics on an individual's visit. We used this information to derive a feature space for a Random Forest Regression Model.

In total we run 23 regression models for states with transactions n > 500. Since n = 500 is not particularly large, 90% of the data is used for fitting and 10% of the data is used for testing and validation. It is still important to note, however, that many states had large values of n such as California which has n ≅ 19,000 or Florida with n ≅ 6700.

There are several reasons why we chose a Random Forest Regression. For one we had very large feature space (182 total features) of which we don’t have clear insight as to what is important and what is not; throwing variables out based on personal preference is a large bias. Random Forest (RF) is robust against the “curse of dimensionality”, that is it can handle a large number of features without overfitting. This is because RF creates decision trees with a subset of the entire feature space chosen randomly allowing it to preserve the notion of distance with high dimensional data. Also with such high dimensional data, we were unsure of the exact assumptions for the geometry (i.e linear, quadratic, etc) of the data, but RF protects against such assumptions.

Our features space consists of 182 features including: Age, Ads, Shopping Attempts, Car Configuration, Ads Clicked, Car Mileage, Car Model, Lead Actions, etc. We tried decreasing the feature space based on MeanDecreaseGini (via RF) which a measure of relative importance of each features. However, this did not significantly change our performance accuracy and actually decreased accuracy by a several percentage points.

We create a heat map that reflects the predicted price, given the interactions a user will have with Edmunds online. From this heat map we can easily see that pricing across the states are variant or invariant with respect to the other states when certain parameters are shifted. Specifically we see that the mileage and count of leads for a given user greatly changes the predicted buying price. Whereas, total advertisement clicks do not significantly change the predicted price.

Overall, we found out that the same person in different states will lead to different prices for the car bought. This shows that local predictors based on region are much more reliable than global predictors. Edmunds can use such inferencing to make prediction models based on localized geographic pricing for individuals with specific site experiences and interaction.