Capstone Project 1: In-Depth Analysis

By Laura Gascue

## Can Health Plan Options be Predicted Given Market Characteristics?

If you don’t have access to health insurance through your job, a family member or Medicare, you will turn to the Individual Market. And when you try to buy insurance, the first thing you will be asked is your zip code. Yes! Not everyone will have the same options, it depends on where you live. How much are these options depending on your neighborhood characteristics?

As in the previous analysis, we limit the geographic areas to the US in 2018 where the Individual Market exchanges are federally facilitated and we look at plans offered through the exchanges. It is also possible to buy insurance off-exchanges, but the large majority of policies are sold in the exchanges; they guarantee compliant with the ACA minimum requirements, and qualifying individuals can request subsidies. We compute the number of plans being offered through the exchanges in each FIPS county code on those states and use it as a measure of the level of competition in that market.

The next step is to identify key neighborhood measures that may be relevant to the question at hands. One could presume that a larger market will be more attractive to companies, and allow for market segmentation strategies. As companies negotiate with health providers in the area, a larger pull of patients can lead to lower costs. This make the market more attractive to insurers, with potentially higher profits. Population income level in the area can also play a role in the number of plans offered. Areas with higher income can support higher coverage, with higher premiums and potentially higher profits for insurers.

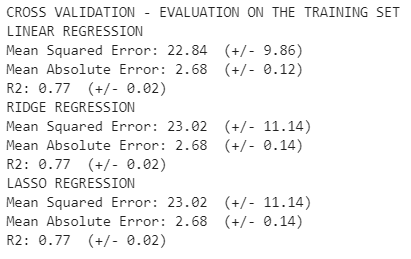
We consider the county’s population size measured by number of individuals living in the area. We also consider the Core-Base Statistical Area (CBSA) classification based on population density, assuming three levels: metropolitan statistical area, micropolitan statistical area, and non statistical area. To take into account population income, we consider county median income level and county unemployment rate. As a measure of providers’ options we include the number of county hospitals. We also add a measure of the individual market size, using the previous year’s individual market enrollment and the number of different insurance companies participating in the market.

We first try a linear regression approach to predict the number of plans faced by consumers in the individual health insurance market and evaluate the explanatory power of these variables. Then we compare the results with a tree regression approach and the random forest. We pay special attention to the number of players in the market.

## Fitting a Linear Model

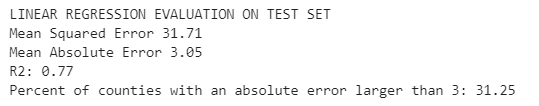
We explore using a model penalizing the number of explanatory variables used, like the Ridge regression or Lasso regression. We split the model into a training and test sets (70%-30%). We use the training set to evaluate the three alternatives using cross validation with 3-folds. Then we use the test set to evaluate the selected model fit.

For a linear regression model we can use the R2, mean squared error (MSE), and mean absolute error (MAE) to evaluate the fit. The R2 shows the proportion of the dependent variable variance explained by the model. When comparing two models, a higher R2 implies the model has better predictive power. The MSE and MAE are measures of the model's lack of predicted power. A lower mean squared or absolute error implies the model has better predictive power. The MSE places a higher penalty to larger errors than the MAE.



When comparing the three linear regression options tested, the scores are very similar (within the 95% confidence interval). The average MSE obtained with the linear fit is slightly better than the Ridge and Lasso regression (though not statistically different), and since the linear regression is widely used, the first inclination is to chose the linear regression model.

An additional test can be run for the selected model using the test set we extracted before. Besides the R2, MSE and MAE, we can also construct an accuracy measure based on an absolute desired maximum error. In this case we are trying to predict the number of plans offered, with an average level of 14.5. We can set the maximum error at 3 plans, about 20% of the mean.



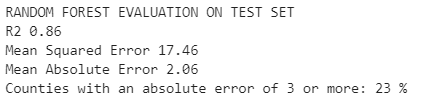
As expected, the fit on the test set is slightly worse than when testing using the training set, even after cross validation. Can this model be improved? We turn to ensemble methods

## Decision Tree Regression and Random Forest

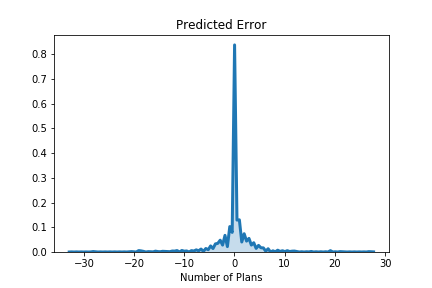
An alternative to the linear regression is to fit a classification tree regression. This approach allows for additional flexibility, taking into account non-linear relationships with the dependent variable. The algorithm tries using each explanatory variable (or feature) to split the target variable in two groups at each node, evaluating the split using the desired score. In this case we use the MSE, penalizing larger errors more than the MAE.

The decision tree regression tends to over fit the sample and get poor results when applied to a data set different from the training one. This method can be strengthen by randomizing the features and samples, and fitting the model multiple times. This is called the Random Forest Regression. By fitting the tree on a subset of the training set and testing on the rest we avoid overfitting. It prepares the tree to fit on unknown values. Once multiple trees are estimated this method uses the average predicted values across all trees.

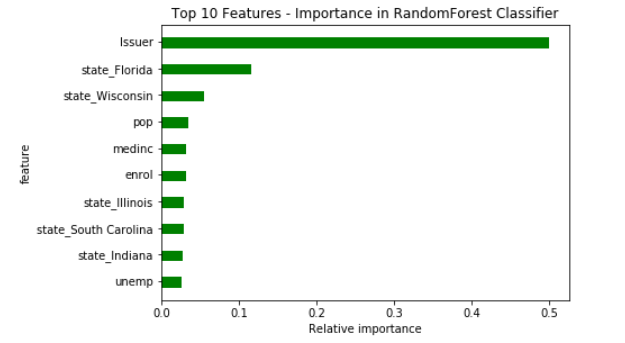
We again split the data into a training and test set, use the training set to fit the forest, and evaluate the fit on the test set. We use 100 samples (trees), with the randomize option but we include all features in the trees. Using a random forest model gets a better fit than the linear regression approach.



There is still a significant unexplained variation in the error term. The bulk of the distribution is within an error of 3 plan options. However, there are still a little over 20% of counties were the predicted error is above this number.



Looking further into how you get to this fit, we observed that there the main results are driven by a couple of features. The number of issuers present in the market is the main feature used to predict the number of plan options. The second two highly relevant features are the indicators for the states of Florida and Wisconsin. These two states seem to behave like outliers when compared with the rest of the states. Population size in the area, county median income, and enrollment in this market in the previous year follow in relevance (see Top 10 Features figure).



## Conclusion

The number of health insurance plans offered in any particular geographic area in the US where the exchanges are federally moderated shows high levels of variability. The number of plans can be predicted using market characteristics with different approaches. The best predictive model identified is the Random Forest, with slight improvement over the linear regression fit.