Capstone Project 1
Spring Board Data Science Career Track
Laura Gascue
April, 2020

Are Individual Health Insurance Markets Competitive? Plan Options Faced by Individuals

Introduction

Since 2014, with the Affordable Care Act (ACA), individuals without access to health care insurance through employment, can purchase one in the Health Insurance Individual Market. The system was created to define a set of minimum coverage standards at affordable cost, aiming to widely expand access to health insurance. The mechanism to achieve this relies on a set of regulations placed on insurers and on market competition. The question we explore here is, are these markets competitive?

We could go over the long academic discussion on how to measure competition levels, or discuss for example the Herfindahl–Hirschman Index (HHI). At the end of the day a consumer want to know if they have options. From the firms point of view, the question becomes, is there space for one more product?

In this article we limit the analysis to the observed final result: how many options an individual faces when choosing health insurance in the Insurance Exchange Marketplace, where plans offered have guarantees in terms of coverage and premiums. First we look at the distribution of number of plans offered, and explore correlations with geographic area characteristics. Second, we consider relevant factors that could be used to explain this measure. We look at state and governors' politic affiliation, population size, unemployment level, number of health professionals or health facilities available in the area, population age and population income, among other variables. We also consider the number of issuers in each area. Finally we try to predict the number of health insurance plans offered in the marketplace for individuals using geographic area characteristics.

The analysis is restricted to the health insurance plans being offered in the US in 2018 through the federally facilitated market place. There are 39 states under the federal program. An individual will have access to the plans offered in the area they live (offered in their county). We compute the number of plans being offered in each FIPS county code area on those states and use it as a measure of the level of competition in that market.

Data Sources

For this project we are combining three different data sources: a) a list of Individual Market Health Insurance Plans, b) socioeconomic indicators at the county level, and c) a table with State Governor's political party affiliations.

The Centers for Medicare & Medicaid Services uploads health insurance plan information on the federally facilitated marketplace through the healthcare.gov website. The latest complete data available corresponds to the plans available in 2018 (the data can be downloaded from 2018 QHP landscape data). The file comes in csv format, with variable names on the first row of the file.

The analysis is limited to the states running under the federal government platform in 2018, and the set of qualified health plans (QHP) offered. Some states, like California with Covered California, chose to run their own Marketplace. The original file has 39,348 rows and a large number of columns. We only read in a subset of columns, containing plan identification variables, plan type, geographic area where offered (at the FIPS County Code Level), and plan premiums, a total of 16 columns. The database cover 39 states: AK, AL, FL, AR, AZ, DE, GA, IL, MI, HI, MO, IA, IN, MS, KS, KY, LA, ME, MT, PA, ND, NJ, NE, NM, NH, NC, OR, WI, NV, OH, OK, SD, SC, VA, TX, TN, UT, WV, WY. The number of total distinct plans offered is 2,722.

The second database we use contains population demographic and health data for the US at the county level. The HRSA Data Warehouse gives access to a dataset with population demographic and health related information through their website (data.HRSA.gov). For this analysis we use the Area Health Resource File (AHRF) database for 2018-2019 (https://data.hrsa.gov/data/download), US Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Workforce, Rockville, MD). This file is available in ASCII and SAS formats. It contains a total of 7,147 variables and 3,230 records.

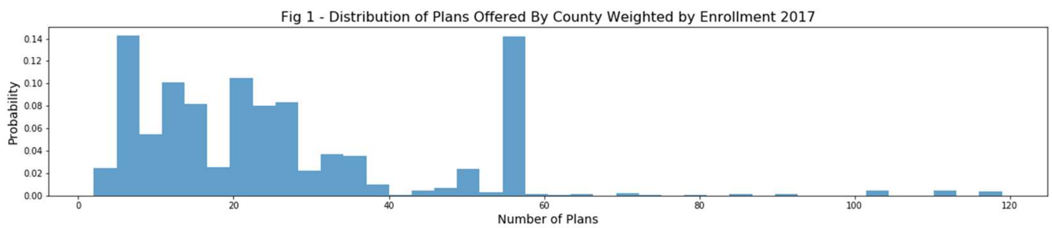
We want to use measures describing each county characteristics in terms of the market size, health services available and income level. The AHRF is a very rich file with more than 7,000 variable. The variables selected are 'Population Estimate 2017', 'Hlth Ins Marketplace Enrollees 2017', 'Median Household Income 2017', 'Unemployment Rate,, 'Total Number of Hospitals'.

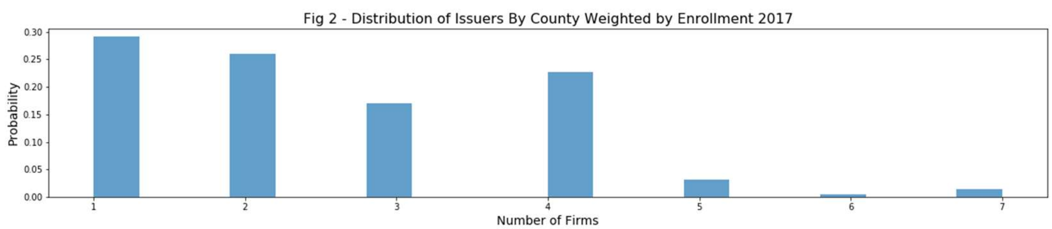
The last source is a csv file with a list of states and their governor party affiliation at the beginning of 2018. This file was compiled through web searches (All_States_Governors_Party.csv).

Plan Options Distribution

The number of plans offered across all counties participating in the federally mandated markets ranges from 2 to 119, with an average number of 14 plans, and a median of 12 plans. The distribution is concentrated in the lower end with an interquartile range of 11. Half of the counties offered a number of plans between 7 and 18. In order to factor the size of each geographic area, and compare with the options faced by all individuals in those areas we use the 2017 enrollment in these markets. Fig 1 shows the distribution of the number of options faced by individuals in this market.

However, many of these plans target different individuals and may be offered by the same insurance company (issuer) applying market segmentation strategies. Taking this into account, we can look at the number of issuers participating in each area. The options now are highly reduced. The number of issuers range from 1 to 7, with 75% of the areas having one or two participants. Fig 2 shows the distribution of number of issuer options faced by individuals in this market.





Since we want to predict the number of options an individual faces when searching for health insurance coverage in the ACA Individual Marketplace we combine the three data sources into one DataFrame. Using the list of plans offered within each county, we create two counts, the number of plans available in that area and the count of distinct insurance companies offering those plans. We use county FIPS codes to identify each geographic area, aggregating the source file to have unique entries for each FIPS code.

The AHRF is read using a data dictionary with variable names, labels, and position. We select a subset of entries from the dictionary, with descriptions for a selection of variables: FIPS, CBSA name, CBSA type and status indicators, state, number of hospitals in the area, population, number of enrollees in the individual market, unemployment rate, and median income. All variables area measured in 2017. Numeric variables need to be converted after the data frame is created, and the unemployment variable is adjusted to represent a percentage.

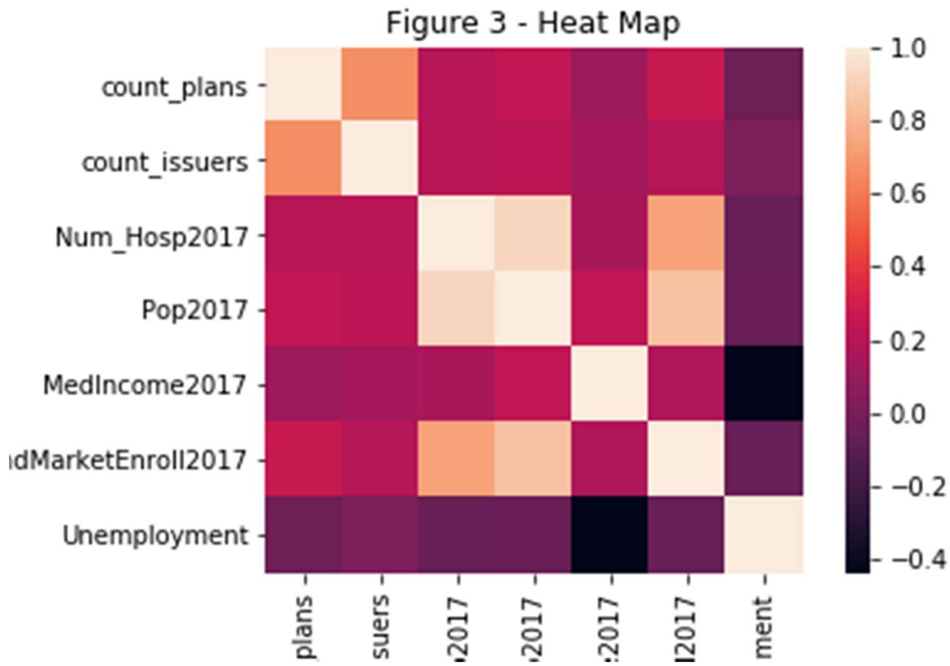
TABLE 1 - Selected Measures

	fips	count_plans	count_issuers	CBSA_Ind	Num_Hosp2017
count	2718.000000	2718.000000	2718.000000	2718.000000	2718.000000
mean	31341.718543	14.461737	1.710817	0.784032	1.804268
std	14959.243238	10.564618	1.004719	0.769837	3.653862
min	1001.000000	2.000000	1.000000	0.000000	0.000000
25%	19135.500000	7.000000	1.000000	0.000000	1.000000
50%	31014.000000	12.000000	1.000000	1.000000	1.000000
75%	46092.500000	18.000000	2.000000	1.000000	2.000000
max	56045.000000	119.000000	7.000000	2.000000	75.000000

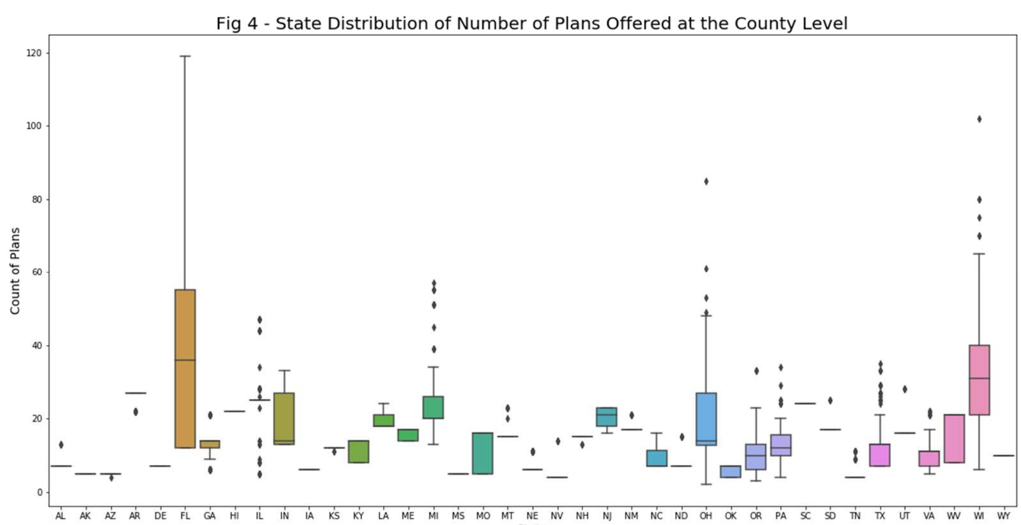
	Pop2017	MedIncome2017	IndMarketEnroll2017	Unemployment
count	2.718000e+03	2718.000000	2718.000000	2718.000000
mean	8.356875e+04	49715.470935	3385.483076	4.637013
std	2.429608e+05	12603.786716	13420.473727	1.633442
min	2.960000e+02	22679.000000	0.000000	1.700000
25%	1.063225e+04	41527.500000	326.000000	3.600000
50%	2.414050e+04	47727.000000	761.500000	4.400000
75%	6.092350e+04	55520.250000	1982.250000	5.400000
max	5.211263e+06	136191.000000	387848.000000	18.300000

One could presume that as there are potentially more enrollees in an area, that market will be more attractive to health insurance companies. 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 look at scatter plots and correlations between the number of plans offered and number of insurers operating in each geographic area in 2018, and the total population in the area in 2017, the total enrollment in the Health Insurance Market in 2017, the unemployment rate. There seems to be a small positive correlation between the number of plans and market enrollment in the previous year (0.27), and the number of plans and population size (0.25).



The distribution of the number of plans available varies widely across states (see Fig 3). Median number of plans offered by state ranges from 4 to 36. Among states offering larger options we see Florida and Wisconsin. In the lower end we see Alaska, Arizona, Mississippi and Iowa. This results suggests that even under federally mandated states, probably state regulations have an effect in these markets.



Another aspect that may help identify areas with different options is population size. We group counties into Metropolitan Statistical Areas (area with one or more urban areaof 50,000 or more), Micropolitan Statistical Areas (areas with one urban cluster of 10,000 people but less than 50,000) and Non-Statistical Area, with no urban areas of 10,000 people or more. The distribution under the two lowest population categories shows areas with similar health plan options and similar dispersion, while the areas with higher population present a tendency to have more health plan options and higher variability.

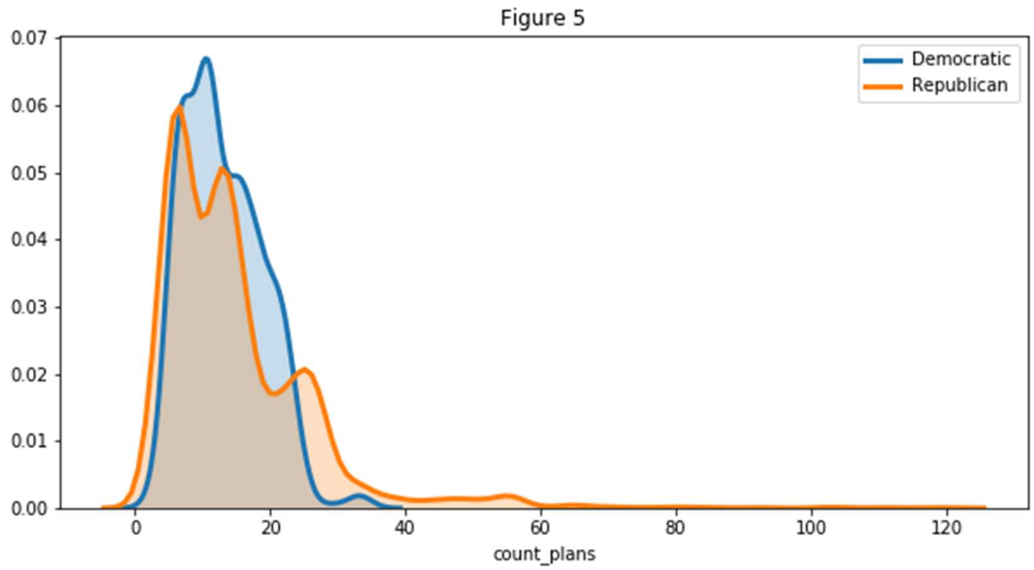
How Number of Plans Offered Changes Depending on Governor's Party Affiliation?

We identify the governor's political party at the beginning of 2018 for each of the states with federally facilitated exchanges under the ACA. There are 9 states under a democratic governor, 29 states under a republican governor, and one state with an independent governor. Note that the number of Democratic states opting to manage their own exchange and not included in this analysis, is proportionally larger than the number of Republican states choosing this option (7 Democratic states vs 4 Republican states). Since we have just one state with an independent governor, we leave this state out of the analysis.

Table 2 shows the distribution of the number of issuers at the county level by state's Governor political affiliation. The large majority of the counties offer health insurance through no more than four companies. States with a Republican governor tend to have fewer issuers with 87% of counties presenting just one or two issuers but a longer right tail. In contrast, counties under a Democratic governor presenting one or two issuers account for 64.5% of all counties. The differences in the distribution of the number of plans is less clear (see Figure 4) but still shows a more concentrated distribution in Democratic counties, but slightly shifted to the right, respect to the distribution in Republican counties.

Table 2 – Federally Regulated States

Number of Issuers	Democratic (%)	Republican (%)	Independen t (%)
1	46.1%	57.8%	100.0%
2	18.4%	29.2%	n/a
3	30.4%	4.1%	n/a
4	4.5%	6.8%	n/a
5	0.6%	1.7%	n/a
6	n/a	0.2%	n/a
7	n/a	0.1%	n/a

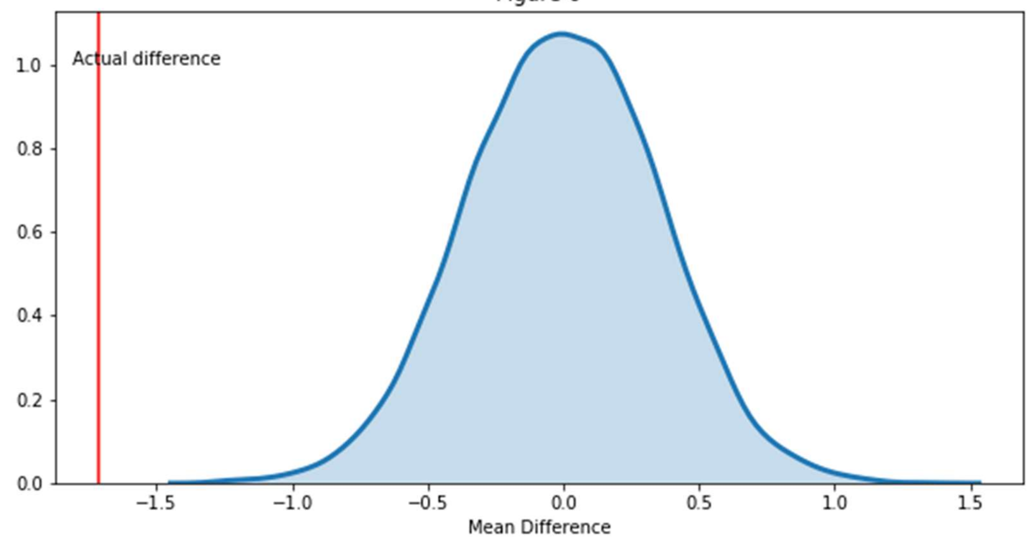


We are going to test the hypothesis that the expected number of plans in states with a Democratic Governor (μ_D) is the same as the expected number of plans in states with a Republican Governor (μ_R).

H0: $\mu_D = \mu_R$

We use a bootstrap approach for this test. Under the null hypothesis the expected value of both means is the same, but their variance is not necessary the same. To generate bootstrap replicates we shift both distributions so their sample means are the same and we sample multiple times from it. We use $\tau = \mu_D - \mu_R$ to test the hypothesis.
We take 10,000 samples from the shifted distribution of plans of the same size as the original data, and for each simulation we compute the observed τ as the difference between both averages. We use the obtained distribution to compute the p-value of the actual observed difference (-1.71). Under the null hypothesis, observing a difference of -1.71 is highly unusual with a p-value < 0.001. Given that a state opts to have their individual health insurance exchange facilitated by the Federal Government, States with Republican governors tend to have more plan options for individuals.

Figure 6



There is a significant difference between counties under a Democratic and a Republican governor in the number of health plan options and individual faces when purchasing insurance through federally facilitated individual market exchanges. The number of options is higher for Republican areas though the size of the difference is small (1.7 in a market with on average about 14 different plans). However, the expected number of insurance companies operating in those counties is lower in Republican areas.

CAN WE PREDICT THE NUMBER OF PLAN OPTIONS 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 multiple 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.

```
Table 2
CROSS VALIDATION - EVALUATION ON THE TRAINING SET
LINEAR REGRESSION
Mean Squared Error: 22.84  (+/- 9.86)
Mean Absolute Error: 2.68  (+/- 0.12)
R2: 0.77  (+/- 0.02)
RIDGE REGRESSION
Mean Squared Error: 23.02  (+/- 11.14)
Mean Absolute Error: 2.68  (+/- 0.14)
R2: 0.77  (+/- 0.02)
LASSO REGRESSION
Mean Squared Error: 55.75  (+/- 25.28)
Mean Absolute Error: 5.06  (+/- 0.37)
R2: 0.44  (+/- 0.03)
```

When comparing the three linear regression options tested, the Lasso regression clearly doesn't do a good job and the other two have very similar scores (within the 95% confidence interval). The average MSE obtained with the linear fit is slightly better than the Ridge 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.

```
Table 3
LINEAR REGRESSION EVALUATION ON TEST SET
Mean Squared Error 31.71
Mean Absolute Error 3.05
R2: 0.77
Percent of counties with an absolute error larger than 3: 31.25
```

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

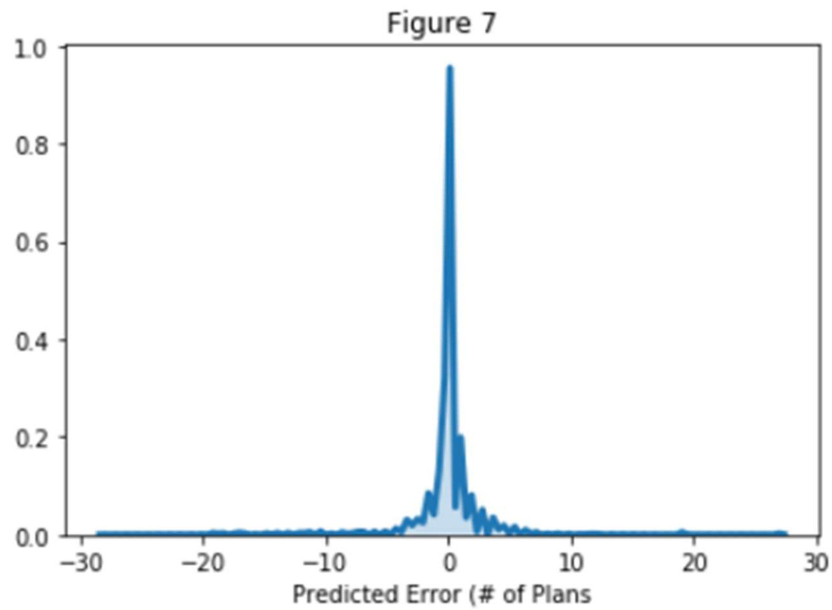
An alternative to the linear regression is to fit a classification tree regression. This approach allows for additional flexibility, taking into account non-linear relationships with the dependent variable. The algorithm takes one explanatory variable (or feature) at a time and uses it to split the target variable in two groups. These decision points or nodes are defined by 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.

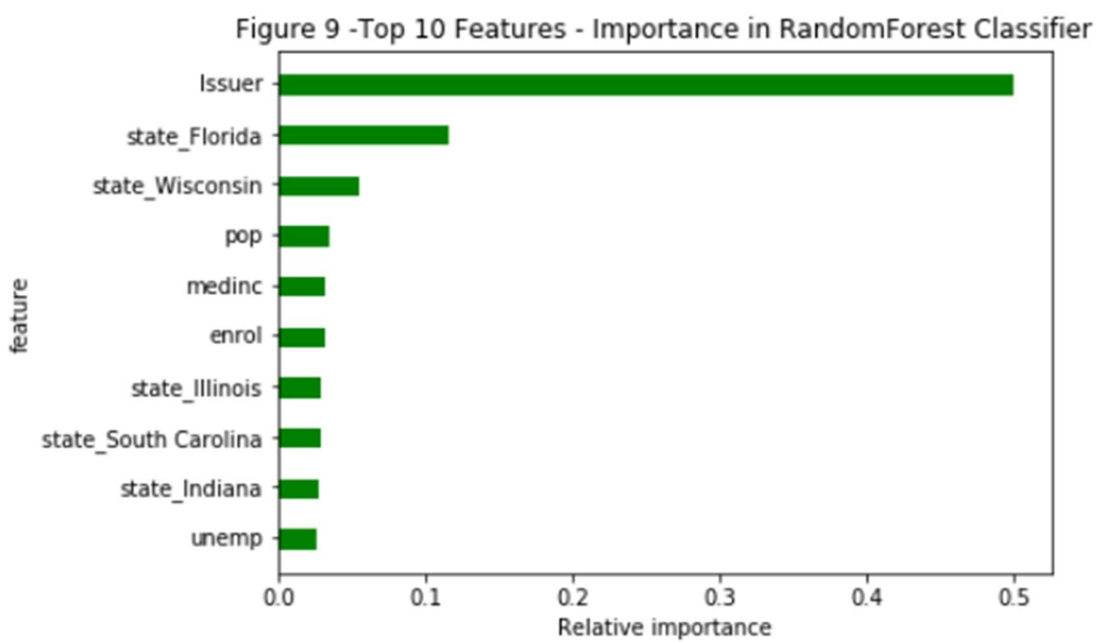
We again split the data into a training and test set, use the training set to fit the forest, and evaluate the fit on the test set. We use 100 samples (trees), with the randomize option but we include all features in the trees. Using a random forest model gets a better fit than the linear regression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

```
Table 4
RANDOM FOREST EVALUATION ON TEST SET
R2 0.93
Mean Squared Error 9.92
Mean Absolute Error 1.30
Counties with an absolute error of 3 or more: 11 %
```

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 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 issuers in each market is very low, with many areas limited to just one company. This points to a low level of competition, though due to high variability in the number of plans, some areas still present the consumer with a significant number of options.

The distribution of the number of plans differs by state. Since the ACA implementation and regulation has been a highly political topic, we test whether the state's color (Governor's political party affiliation) has an effect on this distribution. We observe a significant difference in the number of plans offered in Republican states than in Democratic states, though this difference is small in size (1.7).

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. Besides the high correlation between plans and number of issuers, other variables still have additional predictive power. In particular, two states seem to deviate from the rest, Florida and Winsconsin.

Further analysis into the state facilitated markets is necessary to better inform the prediction of plans offered across the US. From this preliminary analysis it is clear that the health insurance market is highly concentrated and the options a consumer faces highly depend on the ability of the market to attract insurers.