
# Springboard Data Science Career Track Program
# Capstone Project #1

*Author:* *Ashish Mohan Sharma*
*Reviewer:* *Srdjan Santic*
*Publish Date:* *02/27/2019*

# INTRODUCTION

- Health insurance marketplaces, also called health exchanges, are organizations in each state through which people can purchase health insurance.
- A service that helps people shop for and enroll in affordable health insurance.
- The federal government operates the Marketplace, available at HealthCare.gov, for most states. Some states run their own Marketplaces.
- Small businesses can use the Small Business Health Options Program (SHOP) Marketplace to provide health insurance for their employees.
- ACA(Affordable Care Act) or Obamacare, health exchanges were fully certified and operational by January 1, 2014, under federal law.
- Since the inception of the exchange, in year 2014, ACA has been the hot topic of discussion across the U.S.

# Problem Statement

- Many type of plans are offered in various states of United States.
- Each Plan is attached to different Benefits and Limitations
- According to the plan and their associated benefits, monthly premium price is tagged to each plan.
- Nightmare for the member to choose and buy right kind of plan which can serve more benefits with less premium.
- My work will help in understanding the trend of monthly premium in different states and assist in knowing which states are offering best plans with optimum premium

# Potential Clientele

- Companies selling health insurance in state based marketplace or federal run exchange or private plan outside the marketplace.
- Few examples: Freedom Life Insurance Company of America, CIGNA Health and Life Insurance Company, Celtic Insurance Company, Blue Cross and Blue Shield, Aetna Life Insurance Company.
- It may also help the members for future budgeting of their health cost.

# DATA ACQUISITION

- The Health Insurance Exchange Public Use Files (Exchange PUFs) are available for plan years 2014 to 2019 to support timely benefit and rate analysis.
- The link to the datasets are:

https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html

http://www.nber.org/data/cms-marketplace.html

- Files used for the analysis are:
  - **Rate File**
  - **Plan Attributes file**
  - **Benefits and Cost Sharing file**

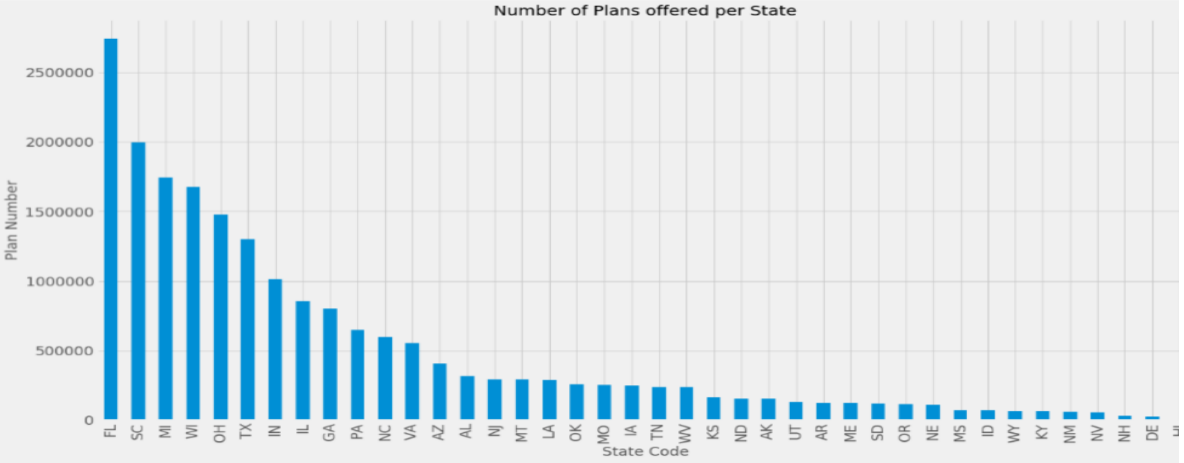
# DATA WRNAGLING AND CLEANING - 1

- Bringing the data to the desired format:
  - Merge the datasets from the different years into 1 dataset.
  - 2017-2019 data columns were different then previous year, so either removed them or changed into desired naming convention
  - Removed columns which were not required to for analysis purpose
- **Imputation**
  - Replace the missing or NAN values with logical imputation
- **Outliers**
  - As the amount of data is huge and very small number of data is outlier, so I deleted the record.
- **Duplicate Rows**
  - Delete the duplicate rows as it may lead to overfitting and will work badly on the unseen data

# DATA WRNAGLING AND CLEANING - 2

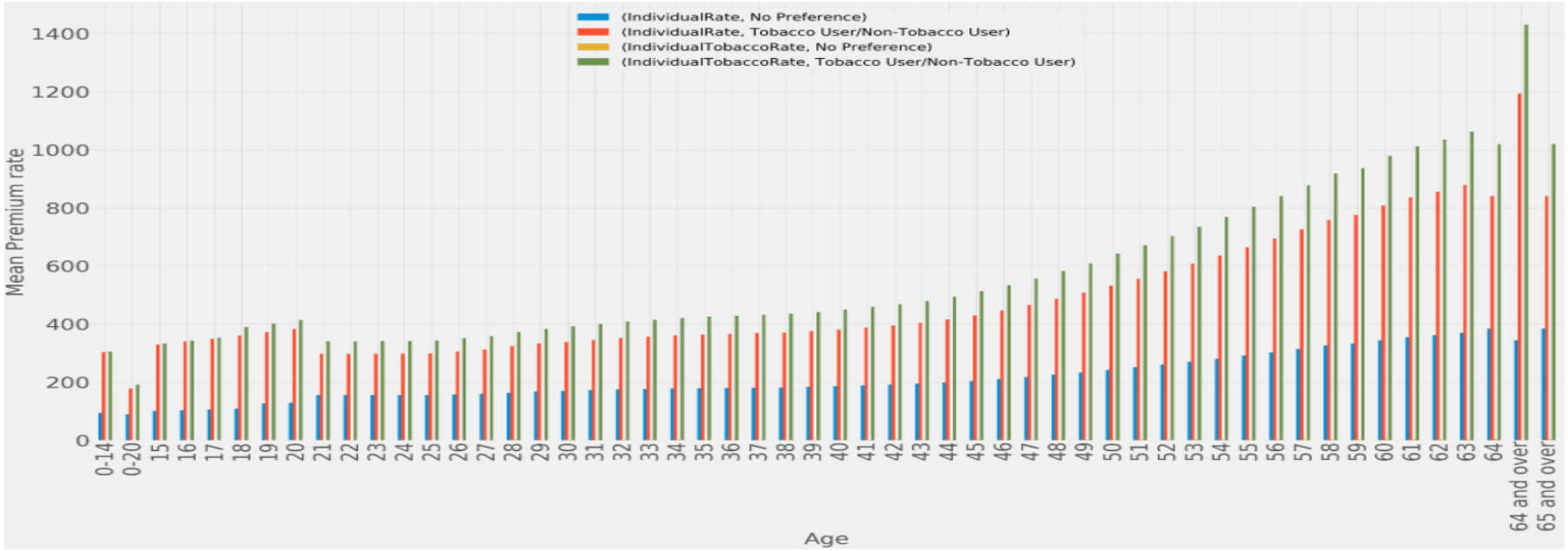
- Columns with “$” sign: to remove the prefix “$”, used replace function and pandas ‘*to_numeric*’ function.
- ‘*Age*’ and ‘*RatingAreaId*’ columns changed from object data type to numeric for prediction model.
- ‘*MetalLevel’* column changed to numeric data type using OneHotEncoder.
- **Feature Engineering**
  - Took the difference of Plan Start Date and Plan End Date and create the new feature ‘*Duration*’
  - Create a new column ‘*Number of benefits*’ by counting the number of benefits offered in the plan
  - Sum up the ‘*IndividualRate*’ and ‘*IndividualTobaccoRate*’, to create the new target variable as ‘*IndividualRateTotal*’
  - Target Variable is changed to Logarithmic as ‘*IndividualRateTotalLog*’

# EDA – Plans offering per state per year

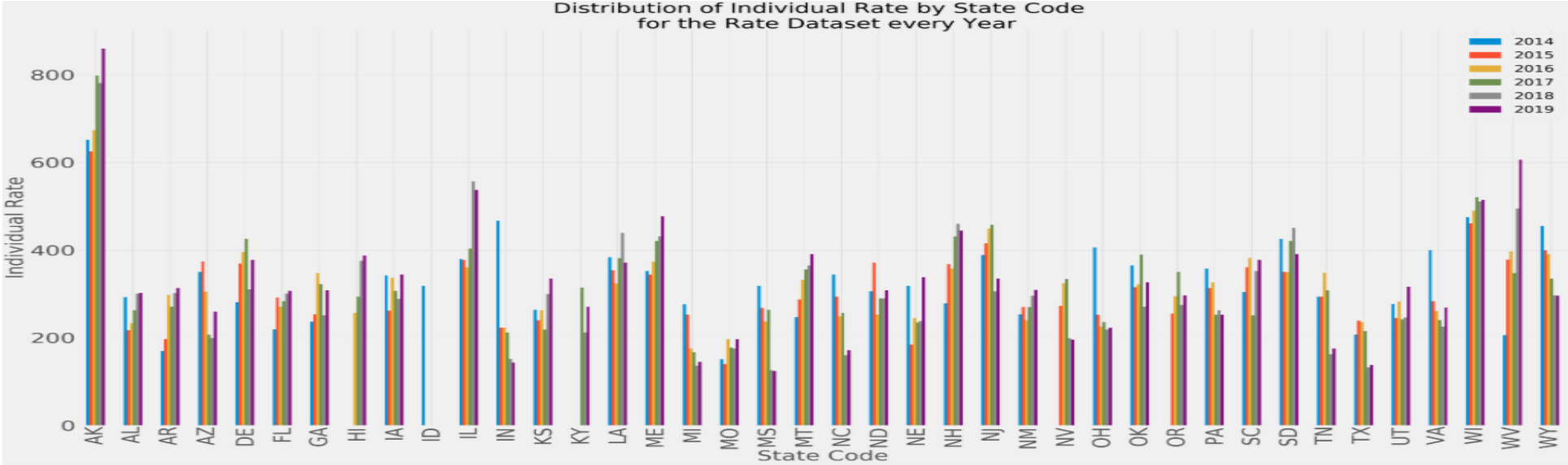

- There is a dramatic change in the number of plans offered in Healthcare market.
- More than 10,000 plans were offered in 2015, reduced to 4,000 in 2019.
- Florida is one of the state which offers the maximum number of Plans to its population
- Alaska and Illinois are not offering that many plans. As these states are sparsely populated so they don't have leverage to provide so much of choice in Plans

# EDA – Premium vs Age vs Tobacco User




- Health Insurance rates go up as a policyholder gets older, with the largest increases coming after age 50.
- Consumers, 64 and older have their premiums capped at 3 times the premiums of the 21 year old base rate.
- There is a constant increase of premium as member ages and it will increase more if the member is a Tobacco user.
- The premium rate for Tobacco user is more than double of Non Tobacco user. This is a good incentive for living a healthy life.

# EDA – Which State is expensive




- Out of 50 states only 40 states have participated in this exchange program
- There are big differences between the states. It's clear that the individual rates in Alaska and Illinois are very high.
- The Median of Alaska is also very high, which means that there are no Plans which comes cheaper in this state
- Tennessee and Texas are the more reasonable states with regards to Individual Plans
- Negative co-relation of Number of plans offered with monthly premium in a State

# EDA – Which State is expensive

**Medical Plans benefits overview**

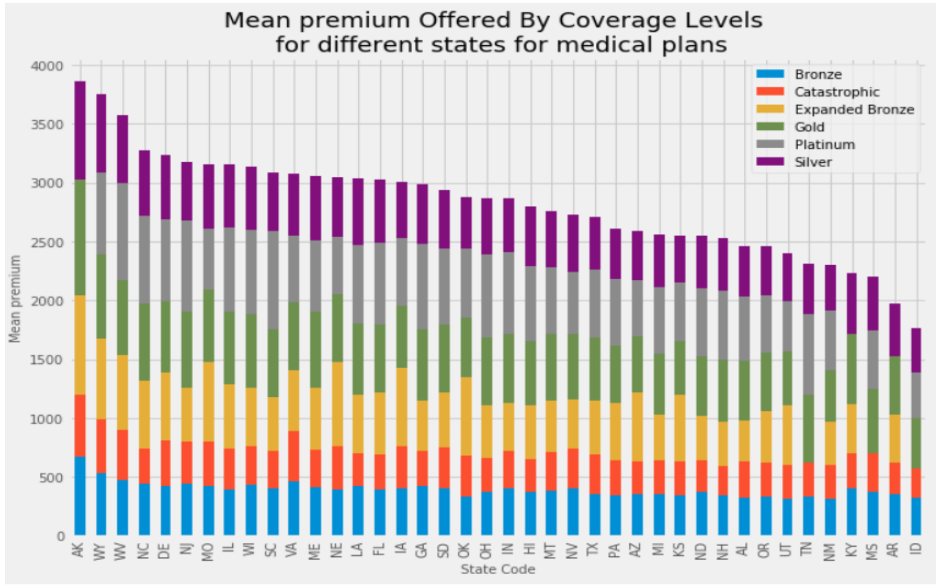
| BusinessYear | Benefits | | | |
|---|---|---|---|---|
| | count | unique | top | freq |
| 2014 | 1134699 | 339 | Mental/Behavioral Health Outpatient Services | 15247 |
| 2015 | 2038625 | 387 | Home Health Care Services | 26991 |
| 2016 | 1737894 | 317 | Orthodontia - Adult | 23138 |
| 2017 | 1298219 | 228 | Accidental Dental | 18568 |
| 2018 | 808437 | 191 | Mental/Behavioral Health Outpatient Services | 11565 |
| 2019 | 947765 | 193 | Routine Dental Services (Adult) | 13618 |

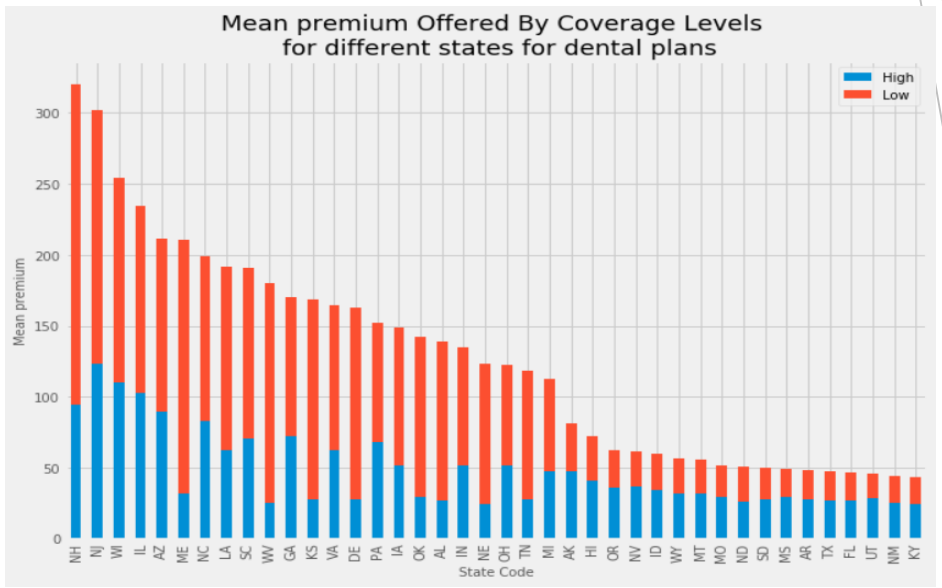
**Dental Plans benefits overview**

| BusinessYear | Benefits | | | |
|---|---|---|---|---|
| | count | unique | top | freq |
| 2014 | 34508 | 172 | Basic Dental Care - Child | 3532 |
| 2015 | 40661 | 141 | Orthodontia - Adult | 4278 |
| 2016 | 36361 | 164 | Orthodontia - Adult | 3859 |
| 2017 | 26056 | 62 | Basic Dental Care - Child | 2803 |
| 2018 | 21215 | 70 | Basic Dental Care - Child | 2292 |
| 2019 | 19285 | 60 | Basic Dental Care - Child | 2077 |

- Few of the benefits are still in the Medical Benefit plan as they are the plans which give both Dental as well as Medical benefits
- Number of Medical plans decreases from the year of inception.
- Maximum number of plans were offered in 2015 but by 2019 it was reduced to 193 from 387, almost 50% reduction.
- Same reduction is in Dental care as well.
- This is not a good sign of the marketplace!!

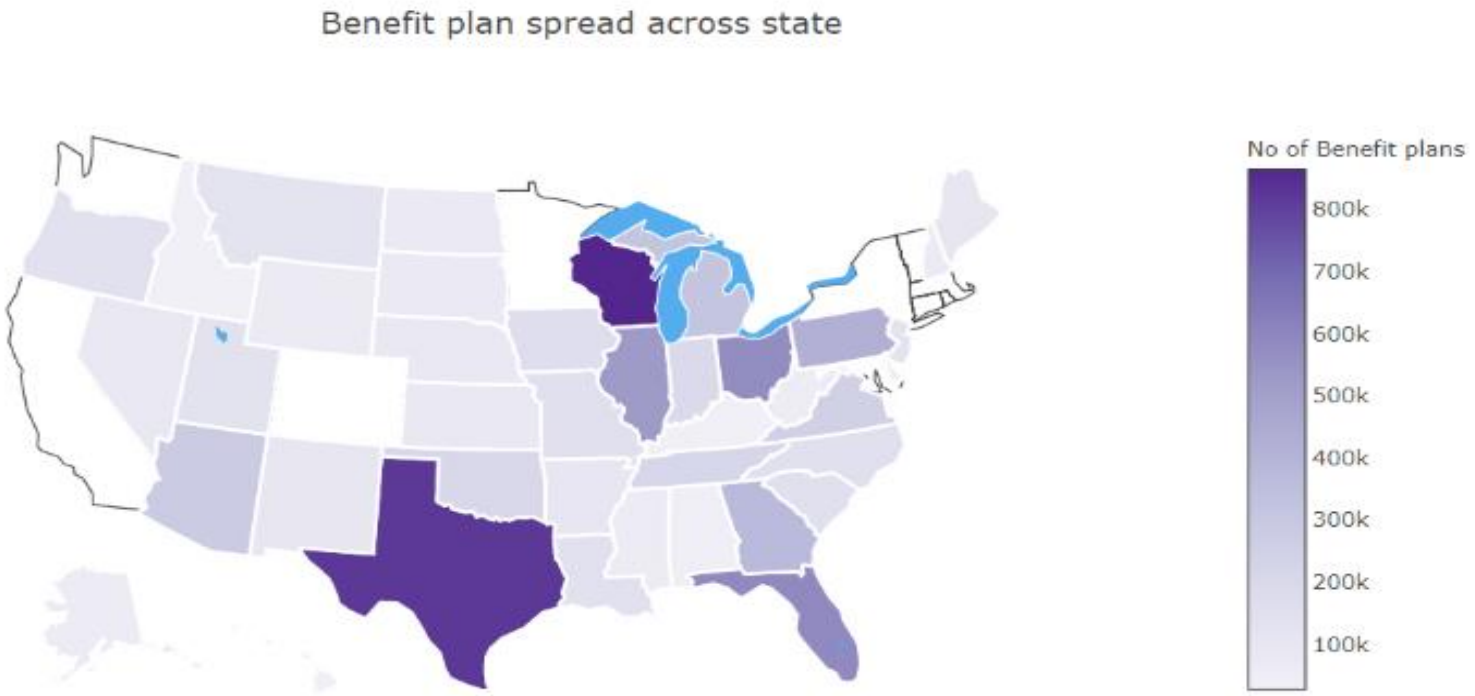
# EDA – Which Plan type is expensive?



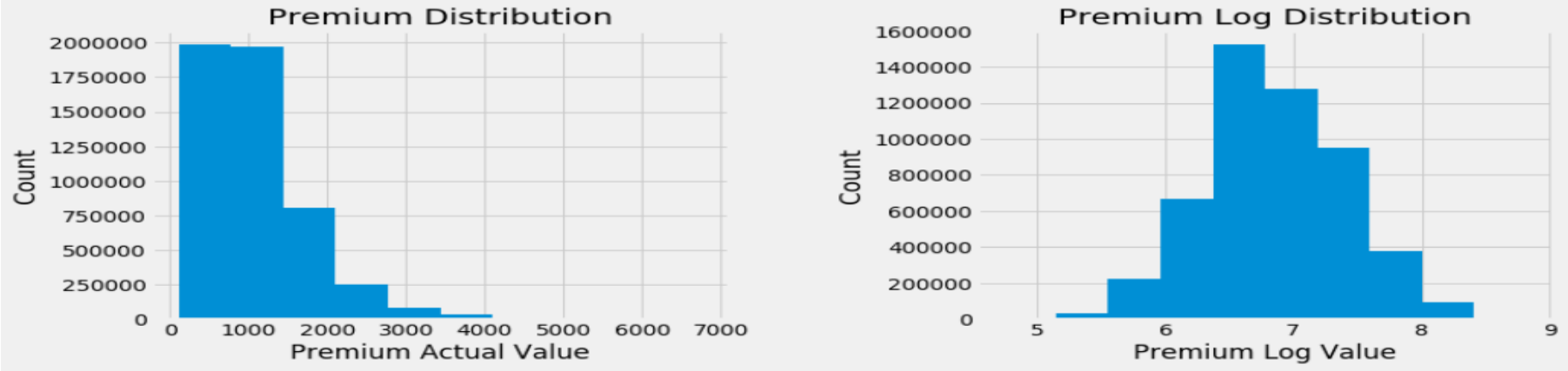

- As expected Platinum plan category is expensive
- Bronze plan catgory is not the least expensive. Catastrophic plans are for specific needs only so they are cheaper but covers few benefits only
- Plan Type ‘Low’ for the dental plan are expensive.
- To our surprise Dental insurance is not expensive in Alaska. In fact it is way cheaper than many states.

# EDA – Benefit Spread Across States




- Wisconsin, Texas and Florida top 3 states to offer highest number of plans.
- In 5 years total of 7.2 Million plan benefits offered to the American people.
- Out of those 7.2 Million plan benefits 964 were unique

# Machine Learning : Data Preprocessing




- Build Regression models to predict the individual premium rates for the Florida State.
- Change '*int64*' to '*int32*', '*float64*' to '*float32*' and object/categorical type of data to '*category*' data type to save space.
- Convert the total premium rate to the '*log*' scale so that its distribution is Normal.
- One Hot Encoding on all my categorical variables.

# ML : Model Training Strategy

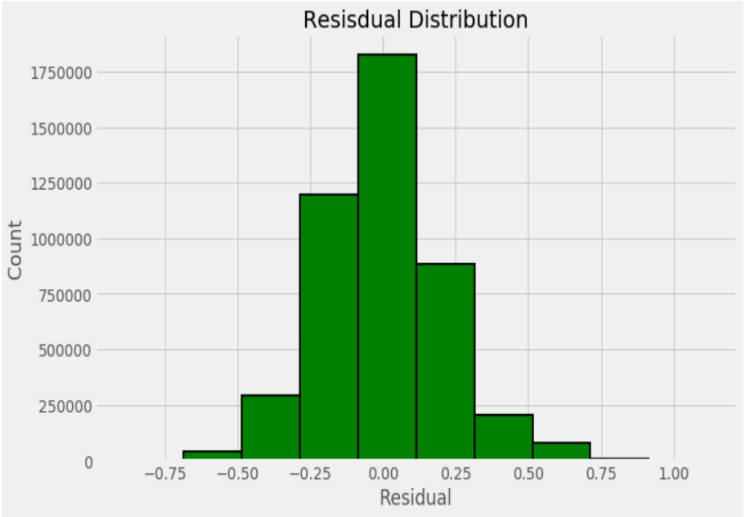
- Training , Validation and Test Dataset
  - 2019 dataset will be used for as Hold Out dataset. Model evaluation will be done on this test dataset
  - 2014-2018 datasets will be used as the training dataset
  - 2014-2018 training dataset will be split into training and validation dataset for hyper parameter tuning or Cross validation purpose
- Models will be evaluated on R-square, Mean Square Error, Mean Absolute Error and Accuracy metrics.
- Baseline Prediction : Predict average of monthly Premium of all the plans from 2014-2018 for the 2019 premium for all plans. Received Negative R-Square value, which suggests it cannot explain any variation in the Test Dataset

# ML : Ordinary Least Squares (OLS)

| OLS Summary Result 1 | | | | | | | |
|---|---|---|---|---|---|---|---|
| Dep. Variable: | Model: | Method: | Date: | Time: | No. Observations: | Df Residuals: | Df Model: |
| IndividualRateTotallog | OLS | Least Squares | Mon, 25 Feb 2019 | 22:08:05 | 4531151 | 4531136 | 14 |
| R-squared: | Adj. R-squared: | F-statistic: | Prob (F-statistic): | Log-Likelihood: | AIC: | BIC: | Covariance Type: |
| 0.848 | 0.848 | 1812000 | 0 | 682010 | -1364000 | -1364000 | nonrobust |

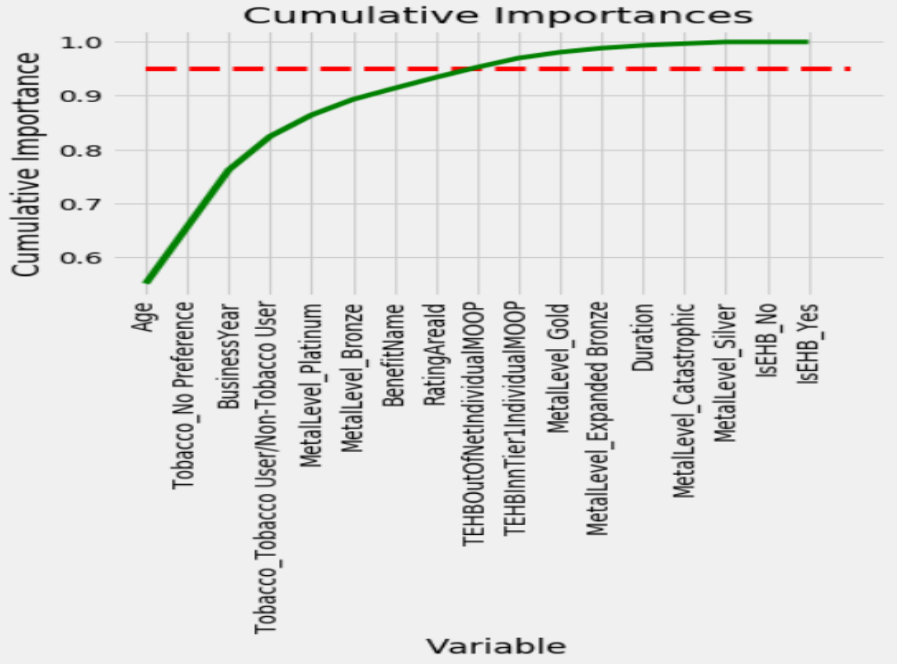
| OLS Test Run Metrics | |
|---|---|
| Mean Absolute Error | 0.166688 |
| Mean Square Error | 0.041974 |
| R square | 0.823046 |

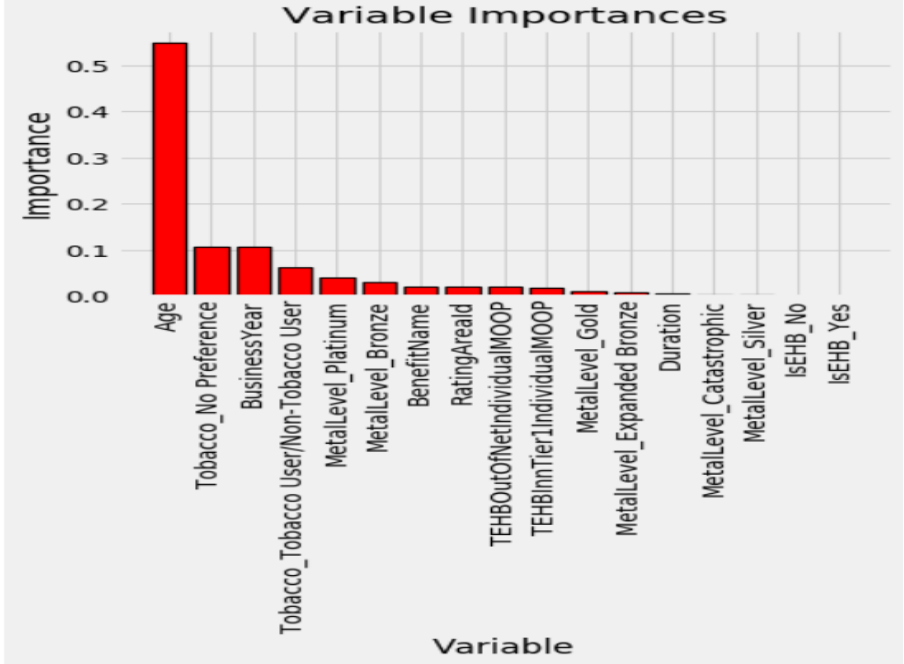
| OLS Train Run Metrics | |
|---|---|
| Mean Absolute Error | 0.159717 |
| Mean Square Error | 0.04333 |
| R square | 0.848488 |




- R-Square value on the training dataset is .848
- Residual Distribution is almost normal with a bit of right skewness.
- R-Square on Hold-Out dataset is .823, which of course better than our Base Model.
- Right skewness suggests that OLS or linear Regression is not the best model to explain the variance. It encourages us to explore other methods now.

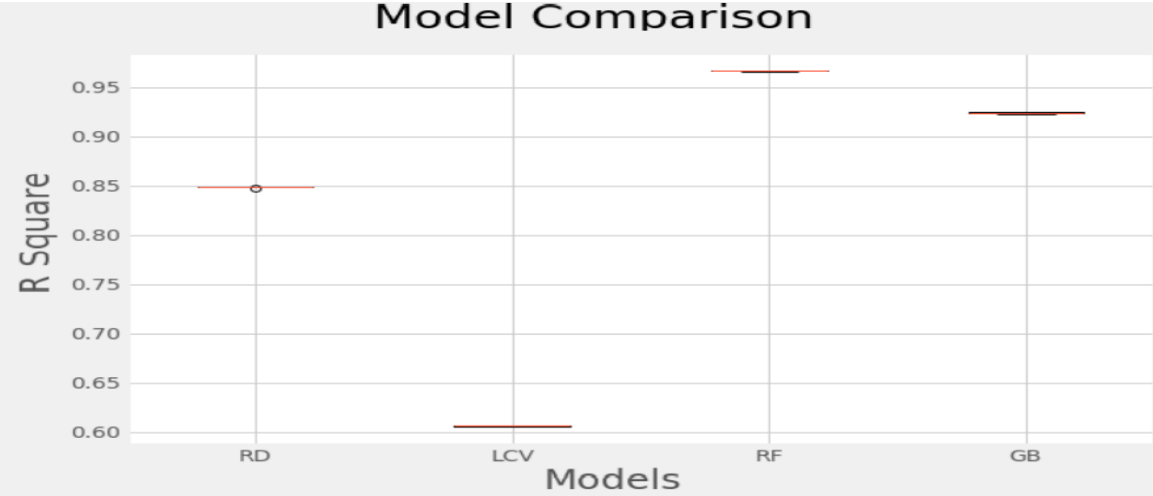
# ML : Decision Tree Regression




- Decision Tree regression helped us to identify the most important features in our dataset
- Cumulative Importance graph suggest that 95% of the variance can be explained by half of the variables in the dataset.
- ‘Age’ alone explains more than 50% of the variation. So it is the most important variable. It also confirms our EDA finding as well about the ‘Age’.

# ML : Model Evaluation





| Model Evaluation result by K-fold | | | | | |
|---|---|---|---|---|---|
| | CV1 | CV2 | CV3 | CV4 | CV5 |
| RD | 0.847774 | 0.848656 | 0.848949 | 0.848799 | 0.848521 |
| LCV | 0.604713 | 0.60523 | 0.606447 | 0.60607 | 0.606141 |
| RF | 0.966181 | 0.966406 | 0.966495 | 0.966441 | 0.966278 |
| GB | 0.924355 | 0.923462 | 0.924485 | 0.923539 | 0.923008 |

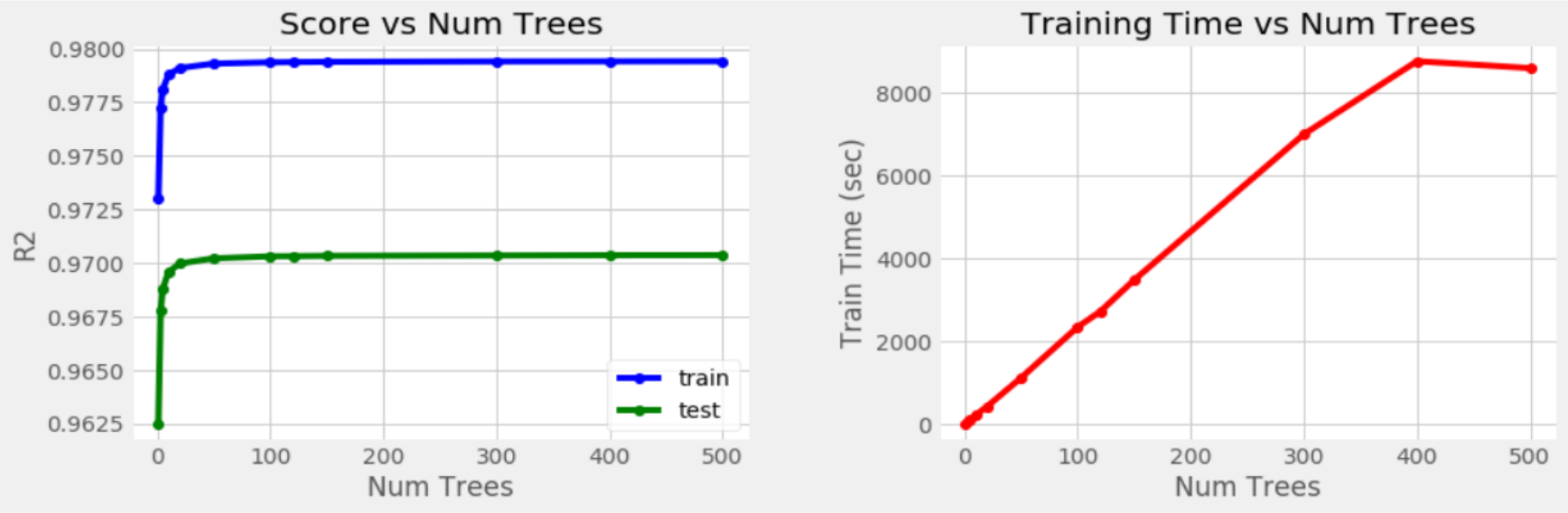
- Evaluate different models on Training dataset using 5-fold cross-validation technique
- Evaluated ‘Ridge (RD)’, ‘LassoCV (LCV)’, ‘Random Forest (RF)’ and ‘Gradient Boost (GB)’
- Random forest produced the best R-Square of 0.966 comparing to other models. So we concentrate on Random Forest to optimize the result and training time.

# ML : Hyperparameter tuning (RF)

| HyperParameters Values for different models | | | |
|---|---|---|---|
| HyperParameters | Base Model Values | Random Search CV Values | Grid Search CV Values |
| bootstrap | TRUE | TRUE | TRUE |
| max_depth | NONE | 50 | 30 |
| max_features | auto | auto | auto |
| min_samples_leaf | 1 | 1 | 1 |
| min_samples_split | 2 | 50 | 10 |
| n_estimators | 10 | 120 | 150 |
| Validation Test Run Results | | | |
| Mean Absolute Error | 0.0634619 | 0.059109038 | 0.059177773 |
| Mean Square Error | 0.0117556 | 0.00838429 | 0.008342143 |
| R square | 0.9588077 | 0.970620925 | 0.970768613 |
| Accuracy | 95.88% | 99.14% | 99.14% |

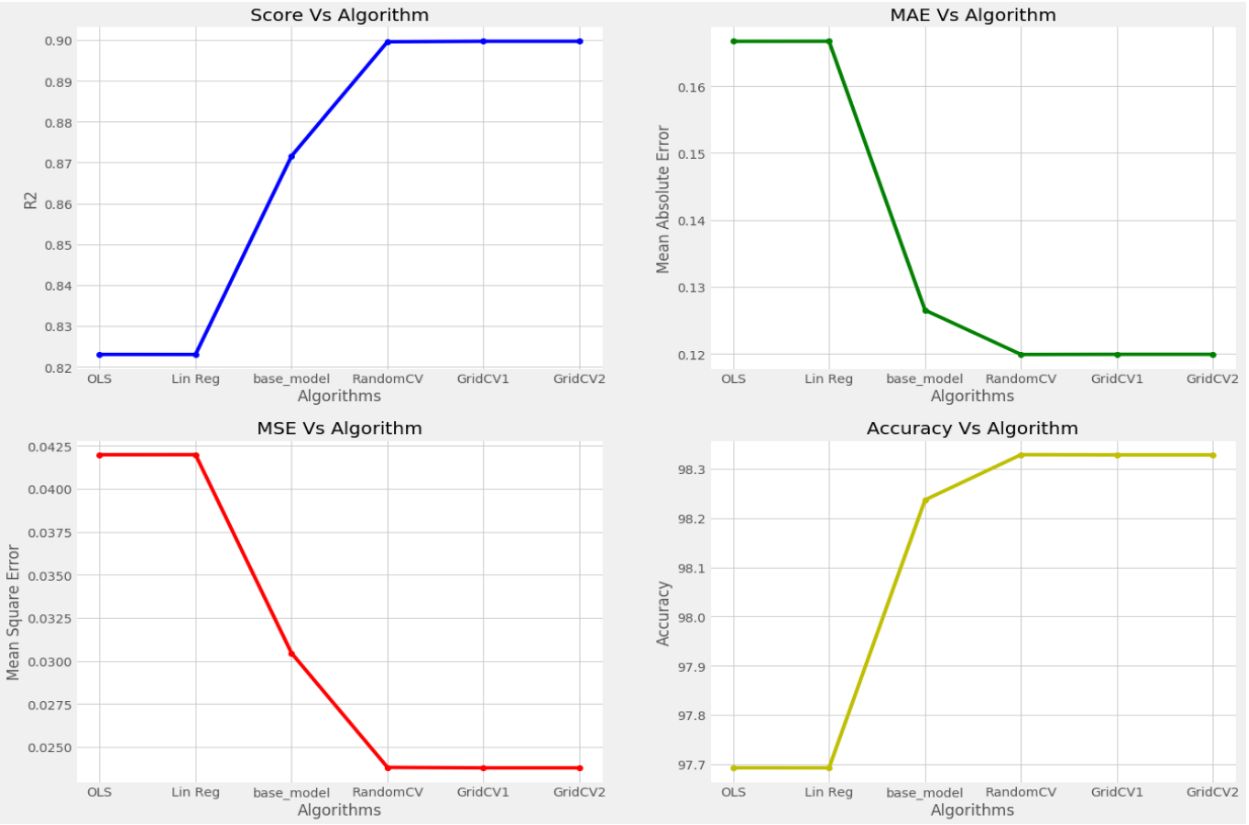
- Split the 2014-2018 dataset into Train and validation dataset.
- Run the base model with n_estimator = 10 and check its performance on our metrics
- Run the Random Search and then Grid Search to tune the Hyperparameters.
- Best performance was received after doing Grid Search CV.
- Marginal difference then Random Search CV results

# ML : Optimizing n_estimamtor




- Kept all other best hyperparameters received from Grid Search CV, and try to tune the n_estimator.
- As n_estimamtor increases there is not much difference in the performance.
- Time to run the model increase steeply as we increase the number of trees.
- Above graph suggest that the optimum number of trees seems to be 120.

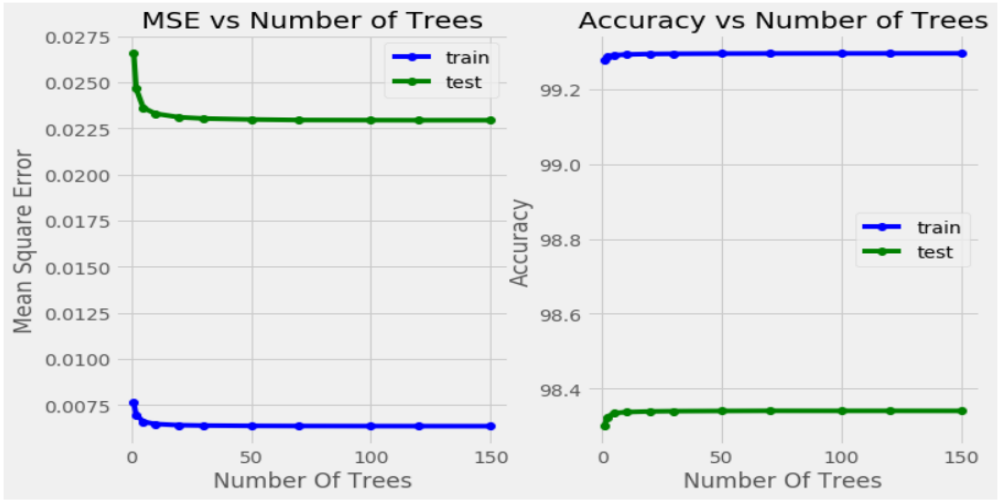
# ML : Performance Evaluation on Hold-Out Data

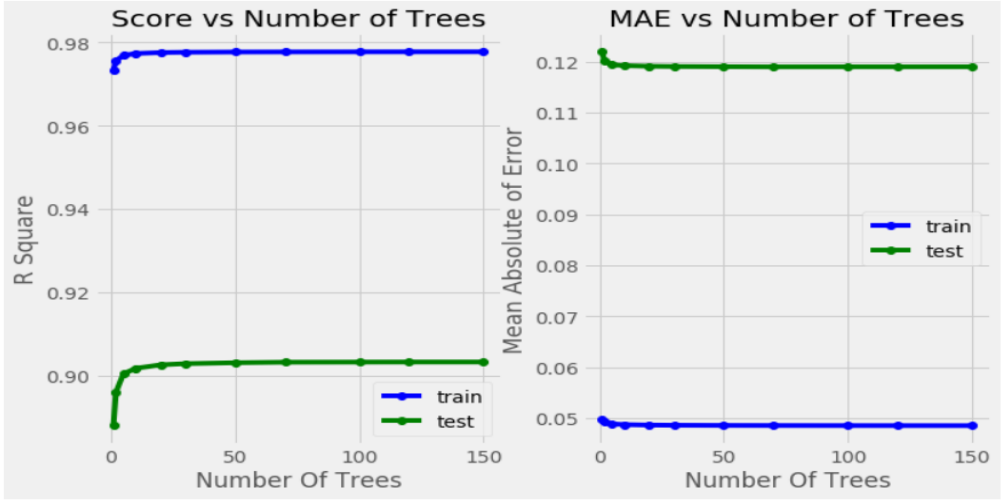



| All Models Performace on Hold Out Data | | | | |
|---|---|---|---|---|
| | R2 | MAE | MSE | Accuracy |
| OLS | 0.823046 | 0.166688 | 0.041974 | 97.69262 |
| Lin Reg | 0.823035 | 0.166697 | 0.041976 | 97.69254 |
| base_model | 0.871604 | 0.126529 | 0.030456 | 98.23696 |
| RandomCV | 0.899587 | 0.119905 | 0.023818 | 98.32891 |
| GridCV1 | 0.899701 | 0.119929 | 0.023791 | 98.32855 |
| GridCV2 | 0.899701 | 0.119929 | 0.023791 | 98.32854 |

- Now run all the models on the Hold Out dataset and check their performances
- Linear Models again did not perform well on the Hold-Out dataset.
- As we search better Hyperparameters for Random Forest, we get better results on Hold-Out dataset.

# ML : Performance on Hold-Out Data Change only n_estimamtor




- Ran the Random forest model with best HyperParameters received from Grid Search CV. Only changing the n_estimamtors from 1 to 150
- This graph helps to trade off the accuracy of the model to expensive computational time.
- There is not much difference in the R-square as we increase n_estimator =100

# ML : Conclusion

- Linear Models are not that accurate in predictions. Though It helped us in identifying the outliers in our dataset.
- Among the Ensemble methods, Random forest is the best model for our dataset
- Tune the hyperparameter for the Random forest using different cross validation dataset
- Random Search CV helped us in narrow down the huge grid of hyperparameters with minimal grid
- Grid Search CV is more stringent and search on all possible combination of the grid provided.
- With Best parameters, we got the accuracy of 98.34% with R-Square as .9037 on our Hold Out dataset i.e. 2019 dataset.
- We may get better result then this but with more computation and time.

# Recommendations & Limitations & Suggestions

**Recommendations:**

- Issuers can target states where the number of plans are less, to expand the business and can provided competitive monthly premium options.
- The model can be scaled to predict the monthly premiums of family groups, like couple with one dependent and couple with two dependent and so on.
- This helps in proper decision making for the consumers before buying heath insurance with their budget and kind of coverage they need.

**Limitations:**

- There are few data points which cannot be shared in public platform because of PHI and HIPPA rules. These data points can help us in getting the better predictions for the premium.
- Due to huge dataset, we limited our modeling to Florida State. Though EDA was done for whole Dataset

**Suggestions:**

- Availability of data about the enrollment in the health insurance market place will help in better analysis, by comparing the actual enrollment with the available options.
- Scale out the model to predict many other premium rates with family group option.
- Dental plan considerations which was not part of this project.

# Thank You!!

Any questions , suggestions or feedback will be appreciated