

## Project McNulty Summary

### I) Design

The project sought to make a binary classify the change in health insurance premia on the Affordable Care Act's Health Exchange into one of two categories, either those with a high increase in prices (defined as 15% or higher) or those without a high increase. While originally a multiple classification was proposed for the project, it was determined that a binary classification would serve more immediate use to those who would want to know about the change in health insurance prices.

The data used came from three major sources, the US Department of Health via Kaggle for health insurance plan data, the US Census Bureau via the American Community Survey for demographic data, and the Robert Wood Johnson (RWJ) Foundation for the county health level indicators. The features used are described in greater detail in part III of this report.

It was decided to apply three filters to the plans to be classified. The first was to look at plans from the year 2015 since that is the latest year for which all relevant data is available. The second was to filter the age to 26 since insurance plans premia for different ages are correlated with each other and since this age group has an additional type of plan relevant for them called a Catastrophic insurance plan. Finally, it was decided to filter the data to only include states where their rating areas and counties were coterminous.

Relevant data about the health insurance plans were extracted from an SQL database found on Kaggle using SQL queries to find the relevant information, including merging rate data from the prior year and the following year. Then, information from the Census Bureau and the RWJ Foundation were merged with the health insurance data to create the table of features to be analyzed for the classification.

### II) Tools

#### A) Python - Programming Language

- 1) Pandas (Data Analysis)
- 2) Numpy (Feature Transformation)
- 3) Scikit-learn (Modeling)
  - a) Random Forest
  - b) Gradient Boosting
  - c) Logistic Regression
  - d) K Nearest Neighbors
  - e) Naïve Bayes
  - f) Support Vector Machine (SVC)
- 4) Matplotlib (Graphic Design)

#### B) SQL - Database Management

#### C) Google Drive / Keynote - Presentation

### III) Data

FEATURE	CONTEXT	SOURCE
Y: over15	Whether a health insurance plan had a rate increase of 15% or greater in the following year	CMS via Kaggle
X <sub>0</sub> -X <sub>4</sub> : Health Plan Coverage Level	The level of coverage provided by plan, dummy variables based on ACA metal level	CMS via Kaggle
X <sub>5</sub> -X <sub>9</sub> : Health Plan Type	The type of health insurance plan	CMS via Kaggle
X <sub>10</sub> : IsNewPlan	Is the health insurance plan new for 2015	CMS via Kaggle
X <sub>11</sub> : IsHSAEligible	Is the health insurance plan eligible for an HSA	CMS via Kaggle
X <sub>12</sub> : OutOfServiceAreaCoverage	Is the health insurance plan cover procedures outside of service area	CMS via Kaggle
X <sub>13</sub> : IsReferralRequiredForSpecialist	Is a referral required for a specialist	CMS via Kaggle
X <sub>14</sub> : NationalNetwork	Is the plan part of a national network	CMS via Kaggle
X <sub>15</sub> : BeginPrimaryCareCostSharingAfterNumberOfVisits	After how many visits does health insurance begin cost sharing	CMS via Kaggle
X <sub>16</sub> : age50plus	The percentage of residents in a county aged 50 or higher	American Community Survey
X <sub>17</sub> : median_age	The median age in a county	American Community Survey
X <sub>18</sub> : adp	The ratio of age-dependent population in a county	American Community Survey
X <sub>19</sub> : % Fair/Poor	The percentage of residents in a county reporting fair or poor health	RWJ
X <sub>20</sub> : Physically Unhealthy Days	The average number of physically unhealthy days reported in a county	RWJ
X <sub>21</sub> : Mentally Unhealthy Days	The average number of mentally unhealthy days reported in a county	RWJ
X <sub>22</sub> : popchange	The percentage population change in a county	American Community Survey

X <sub>23</sub> : % Obese	The percentage of residents in a county that are obese	RWJ
X <sub>24</sub> : % Uninsured	The percentage of residents in a county that are uninsured	RWJ
X <sub>25</sub> : % Unemployed	The percentage of residents in a county that are unemployed	RWJ
X <sub>26</sub> : Violent Crime Rate	The violent crime rate in a county	RWJ
X <sub>27</sub> : Average Daily PM2.5	The average daily level of PM2.5 in the air	RWJ
X <sub>28</sub> : % Severe Housing Problems	The percentage of residents with severe housing problems	RWJ
X <sub>29</sub> : % Long Commute - Drives Alone	The percentage of residents with long commutes that drive alone	RWJ
X <sub>30</sub> : PCP Rate	The rate of Primary Care providers in a county	RWJ
X <sub>31</sub> : Dentist Rate	The rate of Dentists in a county	RWJ
X <sub>32</sub> : MHP Rate	The rate of Mental Health providers in a county	RWJ
X <sub>33</sub> : priorover20	Whether a health insurance plan had a rate of increase of 20% or greater from the previous year	CMS via Kaggle
X <sub>34</sub> : MultistatePlan_2015	Was the health insurance plan a multi-state plan	CMS via Kaggle
X <sub>35</sub> : Age-Adjusted Mortality	The age adjusted mortality rate in a community	RWJ
X <sub>36</sub> : % Frequent Physical Distress	The percentage of residents reporting frequent physical distress	RWJ
X <sub>37</sub> : % Frequent Mental Distress	The percentage of residents reporting frequent mental distress	RWJ
X <sub>38</sub> : % Diabetic	The percentage of residents with diabetes	RWJ
X <sub>39</sub> : HIV Prevalence Rate	The prevalence of HIV in a county	RWJ
X <sub>40</sub> : % Food Insecure	The percentage of residents in a county suffering from food insecurity	RWJ

X <sub>41</sub> : % Limited Access	The percentage of residents in a county with limited access to health care	RWJ
X <sub>42</sub> : Drug Overdose Mortality Rate	The drug overdose mortality rate in a county	RWJ
X <sub>43</sub> : % Insufficient Sleep	The percentage of residents in a county receiving insufficient sleep	RWJ
X <sub>44</sub> : Other PCP Rate	The rate of other Primary Care providers in a county	RWJ
X <sub>45</sub> : Household Income	The median household income in a county	RWJ
X <sub>46</sub> : Segregation Index	The level of segregation between whites and non-whites in a county	RWJ
X <sub>47</sub> : Homicide Rate	The homicide rate in a county	RWJ
X <sub>48</sub> : % Rural	The percentage of residents in a county the live in rural area	RWJ

#### IV) Results

The model that was chosen for the analysis was a Random Forest model, which returns the consensus classification of a group of decision trees. The model showed a high degree of success in predicting whether or not a health insurance plan would have an increase of 15% or higher next year or not. The accuracy score was 97%, meaning that 97% of all plans were predicted correctly. The precision of the model is 91% and the recall is 88%. The AUC of the model is 0.94, meaning that the model has a significant predictive power.

#### V) Next Steps

While the classification did prove to be highly successful, there are several changes that would be ideal to make for the future. The first is to expand the number of years used once the data becomes available in order to account for single-time events that may have an effect on price changes in one year but not another. Another is to expand the geographies analyzed to include other states that do not split their rating areas by county, but rather by ZIP Code or by other methods. A final modification that I would want to make would be to analyze the role of non-Affordable Care Act-compliant plans on compliant plans.