Humana/Mays

2020 Healthcare Analytics Case Competition Written Reports

Team: Imperial Data Force

Team Members:

Xuran Wang

Yuxin Guo

Zhenghao Zhou

10/11/2020

Table of Content

I. Approach

1. Problem Framing

Upon receiving the problem statement, we recognize it as a classic classification problem as the label to be predicted is a binary variable. To tackle a classification problem, we did extensive research on a variety of classification techniques and compared their applicability against our dataset. Random Forest Classifier of Python sklearn is selected by us due to its advantages in the following aspects:

- Easy to handle high dimensionality data
- Robust to outliers
- Quick speed
- Low bias
- Bootstrap class weighting for imbalanced class

The shape of the training dataset indicates a potential dimensionality reduction problem. Additionally, imbalanced class, the distribution of outliers, NAs and "others" tend to cause bias during the removal or imputation of those unspecified or unobserved data. All in all, random forest classification is the approach we believe will bring the best fit to the dataset we had.

2. Data Cleaning and Preparation

In order to clean the training dataset to meet the tidy data standard, we first look at missing values and anomalies by row and column. We conclude that NAs are very likely to be missing-not-at-random, which means pure removal will cause bias. Then we come to the conclusion that imputation with K-nearest neighbours is difficult to realize due to high dimensionality. As a result, imputation with mean for numeric features is applied and one-hot encoding method is applied to deal with NAs and anomalies for object-type features.

Among numeric variables, binary ones were identified and dealt with the way as object variables to retain NAs and outliers. Others are considered continuous values by nature. Additionally, the location-related features, such as state, county and zip-code, are specially treated. They are linked with external data of the 2020 US hospital census.

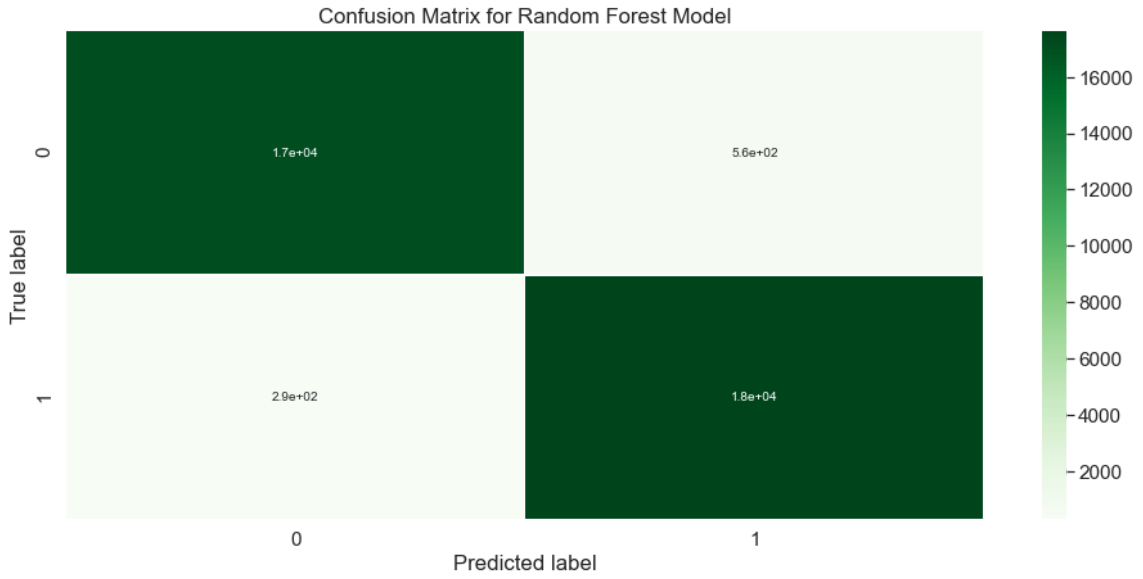
Hospitals throughout the US are grouped by county code. The number of hospitals in each county is calculated and then merged with the training dataset on the county code column, which is formulated by combining the cnty_cd and the state_cd of the training dataset.

We also observe that the training dataset is imbalanced. Only 14.66% of the patients claim to have transportation challenges. Oversampling of the minority class is applied to make full use of the rich data volume. We then use the mean ROC-AUC score across all folds and repeats, confusion matrix, and f1-score to evaluate the performance of the model.

(External data link: https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals)

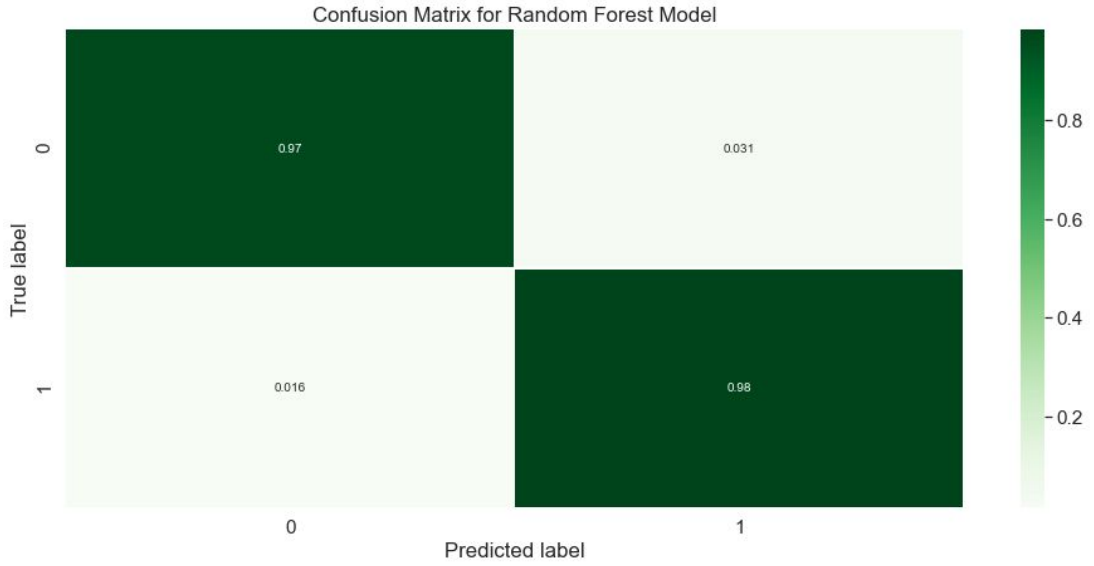
3. Data Modeling

We first use train_test_split from sklearn to divide the training dataset into train dataset and test dataset by 0.7:0.3, of which higher than default test-data size is chosen to mitigate overfitting. The overall accuracy is 97.65%.



- Out of the true values=1, 17482 members are predicted correctly while 238 are incorrect.
- Out of the true values=0, 17305 members are predicted correctly while 600 are incorrect.

- Out of the predicted values=1, 17482 members have true values=1 while 600 have true values=0.
- Out of the predicted values=0, 17305 members have true values=0 while 238 have true values=1.

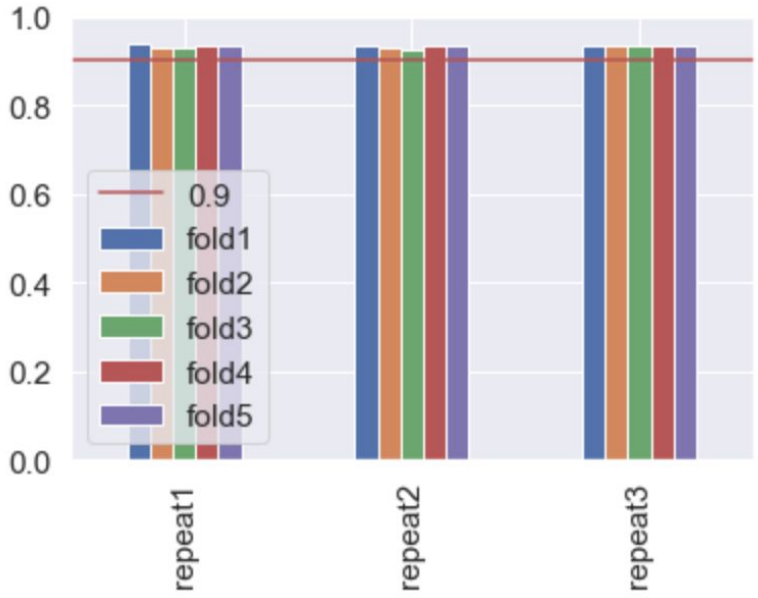


- Out of the true values=1, 98.7% members are predicted correctly while 1.3% are incorrect.
- Out of the true values=0, 96.6% members are predicted correctly while 3.4% are incorrect.
- Out of the predicted values=1, 99% members have true value=1 while 1% have true values=0.
- Out of the predicted values=0, 97% members have true value=0 while 3% have true values=1.

	precision	recall	f1-score	support
0	0.98	0.97	0.98	17698
1	0.97	0.98	0.98	17927
accuracy			0.98	35625
macro avg	0.98	0.98	0.98	35625
weighted avg	0.98	0.98	0.98	35625

The f1-score, precision, and recall are all pretty good. The classification model we trained is robust. We then applied RepeatedStratifiedKFold to cross validate across 5-folds and 3 repeats to evaluate the performance of the model.

ROC-AUC-cross-validation scores of the test set is shown below.



Based on the results of the cross validation, our model is resilient and performs well across different subsets.

Despite our high model performance, we also did hyperparameter tuning.

RandomizedSearchCV is applied to find the best parameter grid.

II. Analytics

1. Tree Plot Analysis

Our team used a tree plot to visualize how the random forest model predicts the value of a target variable by learning simple decision rules inferred from the data features. Every decision at a node is made by classification using a single feature. Plotting a decision tree gives the idea of split value, number of datapoints at every node etc. The insights that are driven from the tree plot are shown below:

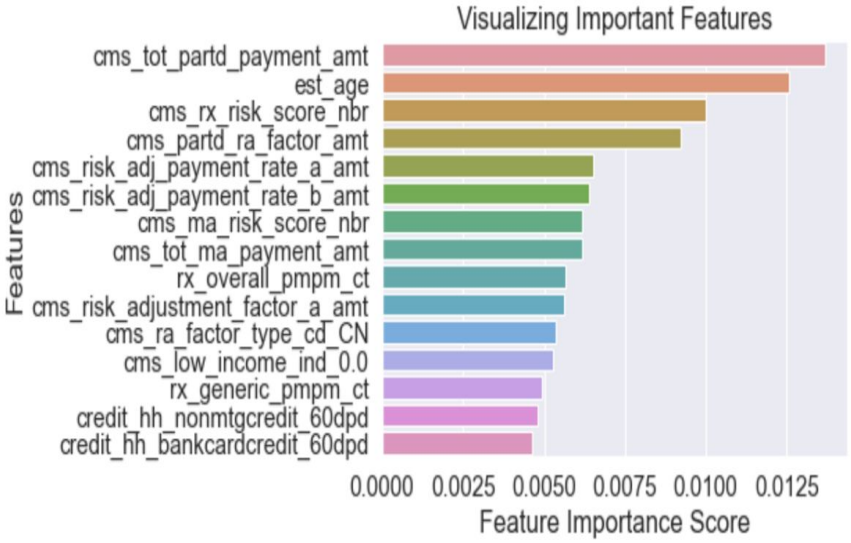
- Humana MAPD members with age less than 64.5 years old are the first threshold to split the dataset into two purity categories. The tree plot splitting paths is

splitting from the criteria whether the Humana MAPD members' age are less than 64.5 years old.

- KBM-Census % Motor Vehicle Ownership and the total Part D Payment Amount are the second and third threshold to determine the tree path.

2. Feature Importance Analysis

One of the most prominent features of Random Forest is the ability to perform feature selection by calculating feature importance scores for each feature. We got rid of the targeted features with zero importance scores to narrow down the analysis scope. We set 0.005 as the cut off point for the feature importance scores as it distinguishes the most number of important features. Twelve variables are selected with cut off point equals to 0.005. We concluded them into five categories. The detailed analysis is shown below.



- Prescription: The total partD payment amount has the greatest impact on the transportation issues. The Medicare partD mainly covers prescription drug costs and individuals need to commute between the pharmacy and home regularly after they receive doctors' prescription. Thus, having reliable transportation would be the first priority for Medicare partD members.
- Age: The second crucial factor for transportation issues would be medicare members' age. Elder members will experience mobility impairment that limits their ability to utilize some "active transportations" such as bicycles and walking. They will also face

challenges while using the local bus system due to its cost, limited routes and hours of operation and safety concerns in certain areas where the bus stops are located. Considering Uber/Lyft and taxis, safety, high costs, and unfriendly drivers could be challenges for these medical members.

- Health Condition: Regarding the Rx Risk Score, which predicts the chronic disease costs of care and mortality rate, this factor would be heavily correlated with transportation issues since chronic disease patients require regular inspection at hospital. The commuting between home and hospital will be extremely challenged.
- Income: The low income variable in the feature selection indicates that medical members' income is related to transportation issues. Individuals who have low income will choose public transportation rather than own a car. In the urban area, it will be an issue because the public transportation will not follow the schedule with complicated road conditions. In the rural area, transportation issues will still exist because the bus stop can be very far from their home.
- Disability: Medical members who have disabilities would struggle in transportation since it would be hard for them to access any kind of transportation.
- Other Financial Issues: The financial issues that members who are experiencing are
- student loans, bank loans, mortgage, and debt, etc. Members who are facing such issues are less likely to spend money on transportations. They tend to rely on public transportation, which could lead to potential commuting problems.
- For a full picture of the Forest, please check this link.

III. Insights

Members are more likely to experience transportation challenges:

1. Have the demand for prescription drugs: members who are enrolled in Medicare Part D, elder members, etc.
2. Elder members: Age above 64.5 years old.
3. Suffer from chronic disease or have poor health: members with higher Rx risk score or higher adjustment payment rate require ongoing medical attention; they tend to have a higher frequency to visit doctors.

4. Low income: members with low income are more likely to encounter transportation problems.
5. Disability: Disabled members are more likely to face transportation issues, they require more accommodations.
6. Have financial issues: Members who have financial issues tend to encounter difficulties in transportation. Financial problems include loans, debt, mortgages, etc.

IV. Recommendations

Based on our previous analysis, our team came up with below recommendations to overcome the transportation barrier for members to access care and achieve their best health.

1. Members with Prescriptions
    - Mail Delivery

      Humana could focus more on on-time mail delivery systems. Such a system will dramatically reduce their visits to the pharmacy, which will minimize transportation risks. Specifically, the company could create a customer survey on their delivery needs that linked to the members account. The member could customize their deliverdy time and frequency based on their preferences and doctor's suggestions.
    - Virtual Doctor Appointment

      Customers could schedule phone calls and video chat with their preferred doctors. By connecting to doctors online, members would save the trip to pharmacies. After diagnosis with the doctor, members could order the prescriptions online and receive them in mail.

2. Senior Members
    - Scheduled Shuttle Bus Service

      Humana could partner with transportation companies, such as uber, lyft, etc. to provide scheduled shuttle bus service. Drivers will pick up members who have appointments and drop them off at the designate hospital. Also, the hospital could

provide special time slots for seniors to schedule an appointment in the early morning, for example, from 8am to 9am since seniors tend to wake up earlier than others. Such arrangements could significantly reduce seniors transportation issues and wait time within the hospital.

3. Members with Chronic Disease
    - Omni Channels Access

      Humana could create omni channels doctor appointments for members, including virtual call, phone call, online chat, etc. These will not only enable more timely, effective communications between members and doctors but also prevent unneeded trips to hospitals. Moreover, we recommend Humana to develop a new feature on the Humana mobile app and website to record members' symptoms on a daily basis. Their personal doctor could have access to these files based on members' consensus. It is easier for doctors to keep track of the members health condition and provide feedback through the app and website. With omni channel access, members will not only have efficient communication with doctors but also reduce the risk of encountering transportation problems.

4. Members with Financial Issues (e.g. Low Income, Debts, Loan, etc.)
    - Endowment

      Humana could start an endowment to provide financial assistance for the members who have financial difficulties. This endowment will allow Humana to invest meaningfully in the communities and help those in need to address income disparities. For example, Humana could provide copay waivers and other proactive actions for certain segments of members who are severely experiencing financial hardship. Members could apply online for financial assistance by providing detailed documentation and ID. Humana Foundation will review these applications and select those who are qualified to receive certain financial aids.

5. Disability
    - Home-based Care Model

      Humana could develop a home-based care model to shift the focus from on site to home based care model. This model will include simple emergency room and

hospital care features to ensure members can handle specific medical problems at home instead of hurrying to the hospital. Home-based care models would avoid secondary infections at hospitals such as flu and COVID-19. Moreover, this model would help Humana to offer high-quality and value-based primary care for patients even at home.

V. Business Value

From the training set, we can see there are 14 percent incorrect predictions of which will not continue to use Humana health service because they have transportation issues. With our prediction model, we are able to reduce the incorrect prediction to 3 percent and the remaining 11 percent of Humana users are likely to stick to Humana health services and possibly to avoid 12 million loss.