Humana-Mays Healthcare Analytics 2020 Case Competition

Transportation Issue Prediction Analysis

Content

The Problem

Transportation Issues on Individual Health and Population Health

The Health and well-being of an individual are intertwined with the social and economic conditions in which an individual lives. The availability of reliable transportation impacts a person's ability to access appropriate and well-coordinated healthcare, purchase nutritious food, and otherwise care for him- or herself. Transportation issues are the third leading cause of missing medical appointments[1] and here is a couple of possible situations one who claims to have transportation issues might encounter;

1. Elisa has a dependable vehicle and has the easiest access to healthcare, but she lives hours away from her primary care providers.
2. Cumberbatch lives in an area where there is an affordable route with a bus stop near his home. The best-case scenario would be him getting to his healthcare provider by taking just one bus. The worst-case scenario could be he might have to switch buses, which can be time-consuming and costly.
3. Julio has neither access to a car or affordable, reliable public transportation. Some people can adapt by walking several miles but he has comorbidities such as asthma that makes him avoid walking on dirt roads near his home.

All possible scenarios might prevent one to delay his or her regular medical appointments and might eventually result in preventable emergency room visits. Other than that, missed appointments cost our country $150 billion annually and medication non-adherence adds another $100 - $289 billion to our healthcare costs each year.[2,3]

Elderly, people with disabilities, low-income individuals and families, veterans, and people with special healthcare needs, which includes people who often need to travel long distances to access care are more likely to need transportation services to maintain their health and well-being.[4] By being able to detect customers who are more likely to have transportation issues can proliferate the number of patients who would make it to their appointments, fill their prescriptions, purchase nutritious food, and take care of daily needs.

The Humana Analytics Competition

Using 1-year worth of retrospective data of Humana MAPD members, competition participants are required to analyze the data to develop a predictive model that takes in a set of features of customers and assigns to each member a probability of being at risk for a transportation challenge. These probabilities will be evaluated by the ROC AUC paradigm, essentially measuring the ability of the model to correctly classify members.

Data Understanding & Preparation

Data Source

Raw training data and the holdout set are provided by Humana. Data comes from various sources documenting members' medical-related behaviors, including:

- Medical claims
- Pharmacy claims
- Lab claims
- Demographic/consumer data
- Credit card
- Clinical condition related
- CMS member data elements

Training data has 69,572 unique records and the holdout set has 17,681 unique records.

Software and Tools

- Interactive Python (3.6.7) in Jupyter Notebooks
- Pandas and Numpy for feature extraction
- Scikit-learn and XGBoost for modeling
- Matplotlib for data visualization
- Microsoft Excel for data exportation

Target Variable

The target variable (transportation issue) is a self-reported binary variable ("0" - No, "1" - Yes) to indicate transportation challenges. Transportation screening question:

"In the past 12 months, has a lack of reliable transportation kept you from medical appointments, meetings, work, or getting things needed for daily living?"

is coming from the Accountable Health Communities - Health-Related Social Needs Screening Tool. The survey was completed in November/December 2019.

Predictive Modeling Approach

Data Exploration

The provided dataset contains 825 features from a variety of sources as already discussed. Key attributes of the data that we will need to consider include:

- **Imbalanced Target Class**: the target variable 'transportation issues' contains 14.7% positive class response (i.e. Transportation Issue = Yes) and 85.3% negative class response (Transportation issue = No)
- **Categorical vs. Numerical features**: there are 22 categorical features and 803 numerical features
- **Binary vs. Continuous numerical features**: there is a mix of binary and continuous numerical features in the dataset, and numerical features have many different ranges and scales.
- **Missing data**: there are 131 features with missing values, ranging from 69,339 missing values in 'hedis_ami' to 228 missing values in 'credit_bal_agency1stmorg_collection'
- **Feature correlation**: many of our features are highly correlated with one another

Model Selection

We initially explored simpler classification models including Logistic Regression, but due to the above attributes of our data, we decided that the tree ensemble model XGBoost would be an appropriate choice to build a predictive model. XGBoost has become extremely popular for machine learning applications due to its high performance across many applications and unique advantages including parallel tree building, tree pruning using a depth-first approach, built-in regularization to avoid overfitting, efficient handling of missing data, and built-in cross validation capability.

The tree-based nature of XGBoost also enables feature importance and feature selection after fitting an initial model, so after basic data cleaning and preparation, we will use the model to identify the most important features to further refine our model.

Data Cleaning

Removing Duplicative Features

Our first step is to remove duplicative features. There are a large number of features that include both a per member per month (pmpm) measure and a binary indicator for the same feature. Since these features are highly correlated, we removed the binary indicator and kept the pmpm measure. We also removed columns that contain all zeros and those with 70% of

more missing values. Removing these columns narrowed down our dataset from 825 to 419 features.

Imputing Missing Values

Our next step was to impute missing values. We used Sklearn Simple Imputer to impute missing values for categorical features using the 'most frequent' strategy and for numerical columns using the 'median' strategy.

Handling Categorical Features

Next, we used Pandas 'get dummies' function to convert categorical features into dummy/indicator variables so that our model can make sense of the information from these features.

Baseline Model

After splitting our cleaned dataset into training, validation, and testing subsets, we were ready to run a baseline model. Our base XGBoost model using default hyperparameters achieved an AUC score of 0.7203, and had very poor precision and recall for the positive class as shown below, meaning that it offered little value in predicting Humana members that are likely to have transportation issues.

	precision	recall	f1-score	support
0	0.869	0.977	0.920	11908
1	0.474	0.123	0.195	2007
accuracy			0.854	13915
macro avg	0.671	0.550	0.558	13915
weighted avg	0.812	0.854	0.815	13915

Even after parameter tuning to optimize this model, we did not see an improvement in these metrics, so we needed to take a different approach to improve the performance of the model.

Final Model

As we explored various modeling options, it became clear that this dataset presented major challenges to building a high performing model. In particular, imbalanced classes, high dimensionality, low correlation between individual features and the target variable, and significant noise preventing a clear decision boundary, required a more advanced modeling approach. After researching modeling approaches to overcome these challenges, we adopted

the following 4 step modeling approach[5]. We believe the below is a strong framework to build upon for this and other noisy imbalanced class problems faced by Humana.

1. **Random sampling with replacement of the majority data**

We use this technique to produce data subsets with random samples of the majority class at a ratio of 1.5x the number of minority class samples to balance the data. Step two will further reduce the number of majority class samples to approach a 1:1 ratio of majority to minority class samples to feed our model.

2. **Tomek link elimination**

Tomek link elimination is a widely used under-sampling technique that identifies 'neighbor pairs' of samples with different class labels. In our case, we apply this technique to remove only the majority class sample of a Tomek Link using Euclidean distance to identify the pairs. This helps to reduce the amount of noise in the data and improves the ability of the model to correctly identify samples in the minority class, which we found to be the major gap in our initial models. After Tomek Link elimination, we are left with class-balanced subsamples to apply XGBoost.

3. **XGBoost classifier fitting**

Our next step is to fit the XGBoost model to each subset of our training data. To prevent overfitting, we tuned hyperparameters using sklearn's CVGridSearch and found that the optimal parameters to maximize ROC-AUC scoring are: learning rate = 0.1, min_child_weight=1, max_depth=4, early_stopping_rounds=10. All other parameters remain XGBoost defaults.

4. **Bagging of individual classifiers for predictions**

Our final step is to take the weighted average of the probability predictions from each individual XGBoost model to reach a final aggregate probability prediction.

Model Performance and Evaluation

Our final model achieved an AUC score of 0.750, which is an improvement from our baseline model but still below what we hoped to achieve.

Figure 1 shows a confusion matrix for our model predictions, and figure 2 shows the classification report. From these two figures, we see a majority class recall of 0.79 and a minority class recall of 0.58, a 21% false positive rate, and a 42% false negative rate.

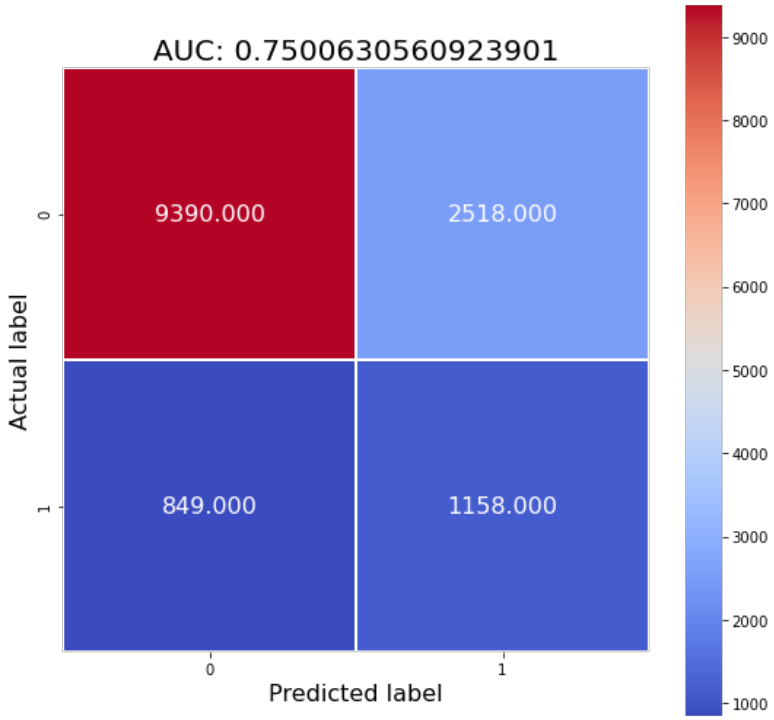

Figure 1: Confusion matrix

	precision	recall	f1-score	support
0	0.917	0.789	0.848	11908
1	0.315	0.577	0.408	2007
accuracy			0.758	13915
macro avg	0.616	0.683	0.628	13915
weighted avg	0.830	0.758	0.784	13915

Figure 2: classification report

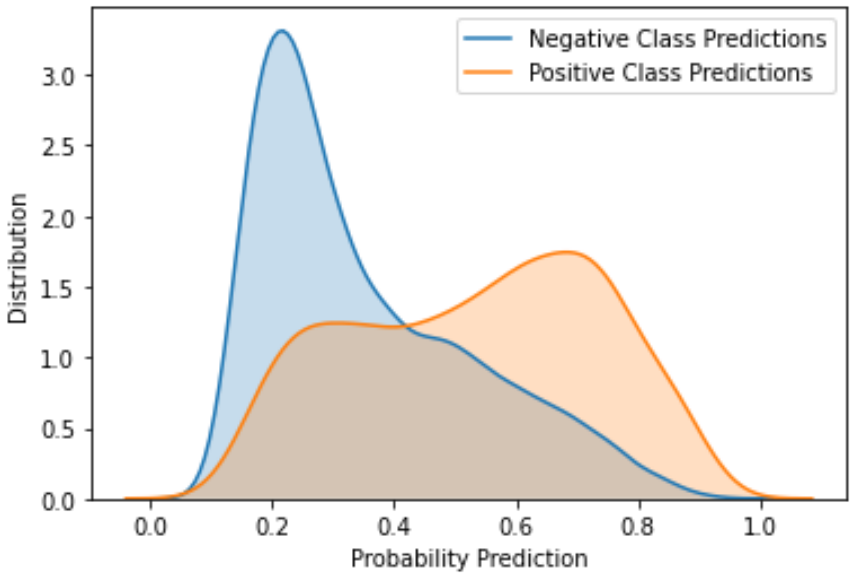

Figure 3: KDE plot of predicted probabilities for positive and negative classes

Figure 3 shows a kernel density estimate (KDE) plot for our predicted probabilities. We see that the model has a strong ability to accurately predict the negative class, but does not predict the positive class correctly at an acceptable rate.

Both false negatives and false positives have a meaningful business impact on Humana, and we include recommended next steps to improve upon this AUC score in the next sections.

Feature Importance and Insights

Figure 4 shows a ranking of feature importances from our model. Key insights from this data include:

1. CMS low-income indicator proved to be the most predictive feature
2. Seven additional CMS features (risk scores/indices/payment rates/payment amounts) appear in the top 25 most important features.
3. Estimated age is highly predictive, but not in the direction that we expected - age is negatively correlated with transportation issues, to our surprise.
4. Some features associated with emergency events proved to be highly predicted
    a. ccsp_239_ind: CCS code for superficial injury/contusion. 28.8% of members with transportation issue had a superficial injury/contusion event compared to 11.5% for members without transportation issues
    b. Total_ambulance_visit_pmpm: total ambulance visits summed from sub-categories. Members with transportation issues have on average 4.5x more ambulance visits (0.13 pmpm vs. 0.03 pmpm).

c. Betos_o1a_pmpm_ct: Per member per month count of logical claims for betos code “ambulance”. Members with transportation issues have on average 4.0x more claims (0.18 vs. 0.045).

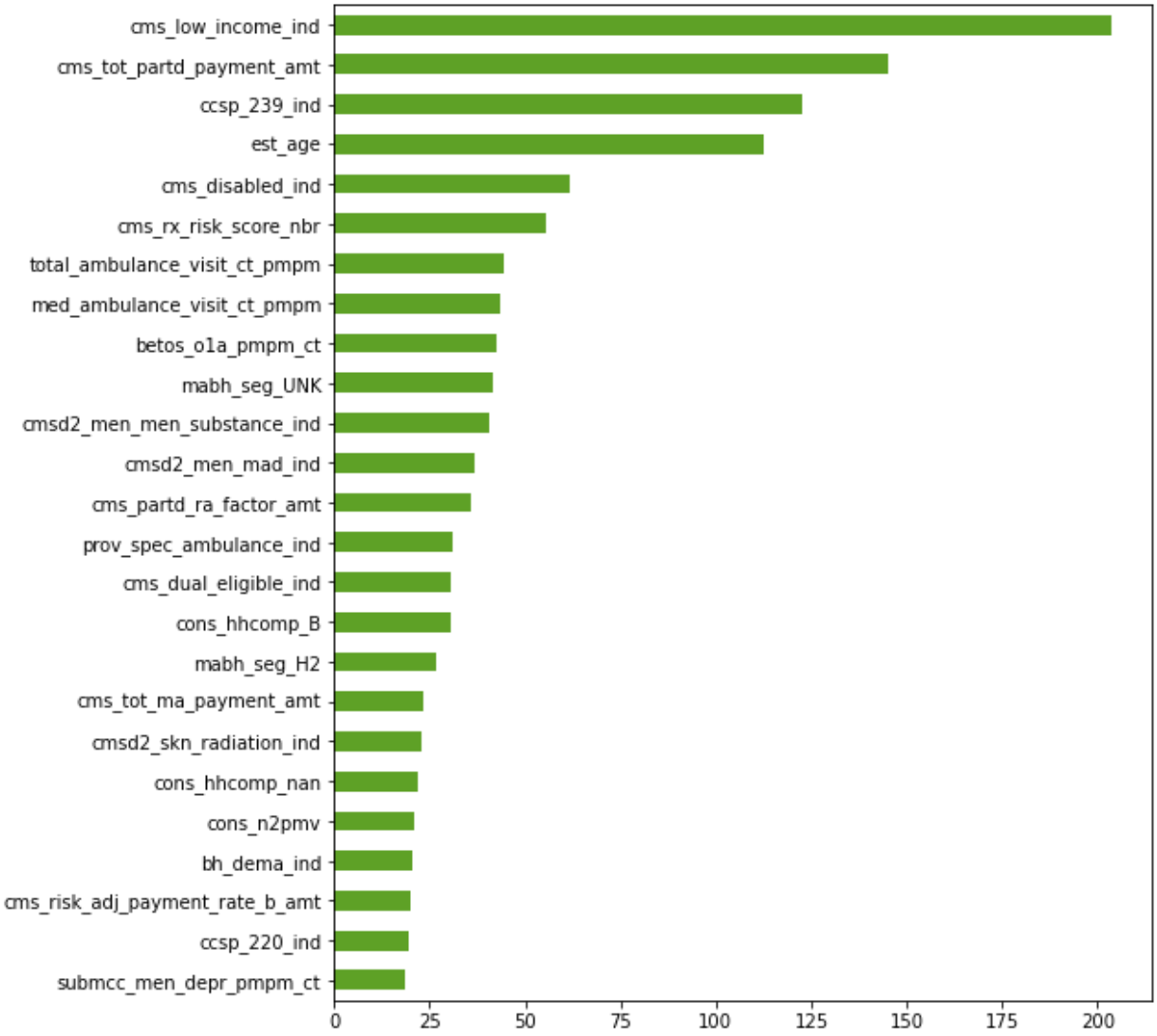

Figure 4: feature importance rankings

Recommendation and Actionability

Recommendation

We first need to establish Key Performance Indicators (KPIs) to measure the business impact. Based on our insights from the most important features we will focus on addressing the increased frequency of ambulance claims for members with transportation issues. Assuming the average ambulance visit cost = $500, and knowing that a member with transportation issues has on average 0.13 ambulance visits per month compared to 0.03 for a member without transportation issues, this equates to a difference of $813 per member per year for this one claim type. We will set a challenging goal to reduce the average ambulance visits per member per month for members with transportation issues from 0.13 to 0.03 over the next 3 years, which will save Humana $813 per member per year based on our assumptions.

Step 1: Segment members with transportation issues into two categories:

1. Disabled shut-ins. This first segment of members would be classified by CMS Disabled Indicator, and we will target at-home health and community solutions towards these members. This group represents 48% of members with transportation issues in our training data.
2. Able bodied with social needs. This segment is characterized most often by CMS low income and dual eligible indicators (41% of members with transportation issues were also flagged low income compared to 20% for members without transportation issues). We will target solutions to enable these members to get transportation to and from medical appointments, grocery stores, work, and community activities.

Step 2: Engage with members with customized solutions

1. Disabled shut-ins. With the specific needs and limitations of this group in mind, we present the following solutions:
    a. Meals on Wheels at home meal delivery program. Building upon a program piloted by the University of Texas Dell Medical School Factor Health program, enroll these members in a program to receive meal delivery from their local Meals on Wheels chapter. Meals on Wheels volunteers deliver food and conducts a basic check-in with the member to assess mental health, basic hygiene, and medication adherence, and feed this information back to Humana to ensure the member is not moving towards increased risk of ambulance/emergency event. Humana should work with Meals on Wheels to develop an outcomes-based payment model to support this program.

    b. At-home primary care program. Connect these members with at-home primary care providers in their area to ensure they are receiving basic preventive care. Incorporate digital services for members with internet/basic technology skills.
2. Able bodied with social needs. These individuals demand a different set of solutions.
    a. Free ride program. Partnering with Uber or Lyft, offer free/discounted rides to and from medical appointments, grocery stores, work, and community events. Tap into programs already in place between Epic-Lyft and Cerner-Uber to enable healthcare providers to order rides for these members.

Actionability

- To improve model performance, we suggest including additional features and feature engineering. Opportunities include more granular geographic features, additional CMS features (since these features proved highly predictive), and more detailed claims data related to ambulance and emergency events.
- Quantify the definition of transportation issues. Definition of having transportation issues can be subjective and vague based on an individual's previous experience and expectations. A better definition would be "within the past 12 months, one has missed 50% of my previous medical appointments, meetings, work, or from getting things needed for daily living.
- Investigate further reimbursement for travel to determine the role it plays in keeping appointments and avoiding fragmented care.
- Objectify outcome measures (transportation issues) such as missed medical appointments, rescheduled appointments, delayed medication fills, and changes in clinical outcomes. This would further clarify both the impact of transportation issues and types of transportation interventions needed.

Reference

1. LaPointe J. 52% of practices use various reminders to stop patient no-shows. RevCycleIntelligence. https://revcycleintelligence.com/news/52-of-practices-use-various-reminders-to-stop-patient-no-shows. Published March 15, 2017. Accessed Oct 2, 2020.
2. Boylan L. The cost of medication non-adherence. National Association of Chain Drug Stores. 2017. https://www.nacds.org/news/the-cost-of-medication-non-adherence/#:~:text=Medication%20non%2Dadherence.,%24100%E2%80%93%24289%20billion%20a%20year. Published April 20, 2017. Accessed Oct 9, 2020.
3. Gier J. Missed appointments cost the U.S. healthcare system $150B each year. https://www.hcinnovationgroup.com/clinical-it/article/13008175/missed-appointments-cost-the-us-healthcare-system-150b-each-year#:~:text=The%20total%20cost%20of%20missed,minutes%20and%20%24200%20on%20average. Published Apr 26, 2017. Accessed Oct 9, 2020.
4. Rural Health Information Hub (RHIhub). Transportation to support rural healthcare. https://www.retrospectivefo.org/topics/transportation. Published April 20, 2020. Accessed Oct 7, 2020.
5. Luo R, Dian S, Chen W, et al. Bagging of Xgboost classifiers with random under-sampling and tomek link for noisy label-imbalanced data. IOP Conference Series: Materials Science and Engineering. 10.1088/1757-899X/428/1/012004.