HUMANA-MAYS HEALTHCARE ANALYTICS 2020 CASE COMPETITION

Social Determinants of Health
Transportation Challenges

Table of Contents

1. Executive Summary

Each year, 3.6 million people in the United States do not obtain medical care due to transportation issues. Transportation issues include lack of vehicle access, inadequate infrastructure, long distances and lengthy times to reach needed services, transportation costs and adverse policies that affect travel. Transportation challenges affect rural and urban communities.

An Integrated Value-based Health Ecosystem is composed of Pharmacy, Behavioral Health and Social Determinants. This study will focus on the social determinants of transportation challenges. Our goal is to develop a binary classification model to predict whether Medicare members are at risk for a Transportation Challenge. We will then provide actionable solutions for members to overcome this barrier depending on the key indicators of the model.

Firstly, we preprocessed the raw data by imputing missing values. Secondly, we performed feature transformation and feature selection to reduce the dimension of the dataset and target significant features. Then, we implemented multiple models to do classification, including Logistic Regression, SVM, Random Forest, XGboost, MLP, and Keras Sequential Model. By training a XGBoost Classifier and validating the model meanwhile, we got a Cross Validation Area Under Curve (AUC) of 0.7538.

We discovered interesting insights and found that most of our top features include low income indicator, disability indicator, superficial injury, estimated age, etc. Based on this, we provide recommendations according to eight main categories of variables. By using the recommendations of our key indicators and model predictions, Humana will be able to propose actionable recommendations to relieve the transportation challenges for Medicare members.

2. Data Understanding and Preprocessing

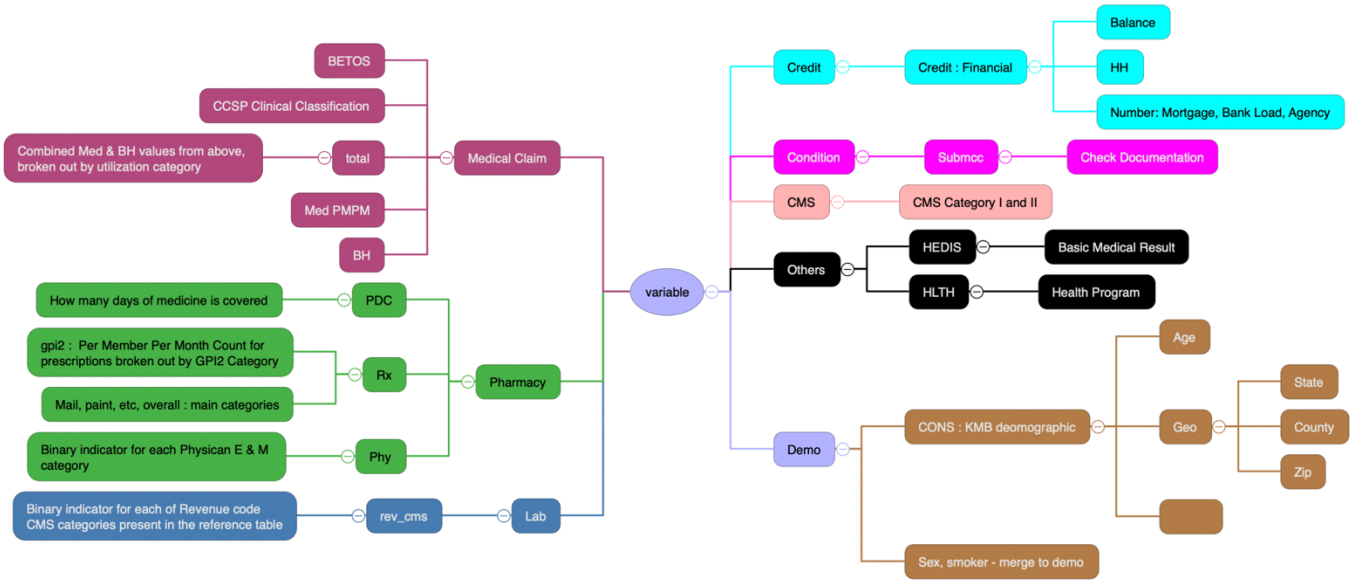
2.1 Data Understanding

We are provided a training dataset of 69572 rows and 824 features except for the unique key person_id_syn and target variable transportation_issue. This provides a 1-year look back for 69572 members before the event collection. There are 14.66% of members facing a transportation challenge. The holdout (test) dataset consists of 17681 rows and 824 features. The ratio of female members to male of the training dataset is 1.444, while 1.415 for the test dataset. The ratio of members with disability to total members for the test data and training data are 30.21% and 30.49%.

Additionally, we compared missing value percentage and value distribution (min, max, mean, median, std, quantile) of each variable in the training dataset and holdout dataset. We found there are many variables having different distributions in two datasets. It reminded us that there could be many outliers in the two datasets and scaling using the mean and variance of the data is likely to not work very well. Therefore, we applied the RobustScaler method to scale data before training.

index	max	mean	std
submcc_cad_mi_ind	0.0%	7.7%	3.9%
submcc_cad_mi_pmpm_ct	40.9%	31.7%	36.8%
submcc_cad_ptca_ind	0.0%	2.3%	1.1%
submcc_cad_ptca_pmpm_ct	33.5%	9.3%	14.0%
submcc_can_brst_ind	0.0%	13.7%	6.6%
submcc_can_brst_pmpm_ct	34.8%	25.5%	33.0%
submcc_can_dig_ind	0.0%	1.0%	0.5%
submcc_can_dig_pmpm_ct	11.1%	43.4%	32.1%
submcc_can_end_ind	0.0%	36.2%	20.1%
submcc_can_end_pmpm_ct	0.0%	20.2%	0.5%
submcc_can_gu_ind	0.0%	4.3%	2.0%
submcc_can_gu_pmpm_ct	20.0%	3.7%	0.7%
submcc_can_h/n_ind	0.0%	59.3%	36.2%
submcc_can_h/n_pmpm_ct	96.7%	88.8%	93.0%
submcc_can_h/o_ind	0.0%	3.5%	1.5%
submcc_can_h/o_pmpm_ct	5.5%	5.3%	16.1%
submcc_can_leuk_ind	0.0%	32.5%	17.9%
submcc_can_leuk_pmpm_ct	75.0%	72.2%	66.1%
submcc_can_lymp_ind	0.0%	25.2%	13.5%
submcc_can_lymp_pmpm_ct	22.2%	17.0%	35.7%
submcc_can_ner_ind	0.0%	68.6%	29.9%
submcc_can_ner_pmpm_ct	239.0%	564.2%	326.4%

The features can be segmented into 8 main categories based on the initials and suggestion from the original dataset documentation:



Among these variables, there are three types of data, float (443), int (360) and object (21).

The object variables are categorical variables, such as cnty_cd (country code), sex_cd (member gender), state_cd (Postal Abbreviation), zip_cd (zipcode), cons_cyms (KBM-Category-Census Education Level). We need to convert these variables into dummy before fitting into the model.

For float variables, since we focus on Random Forest and XGBoost models, we need not to standardize the float variables.

For integer variables, most of the variables are binary indicators, such as for each of the BETOS codes, CCS code, Abnormal Lab Results, prescriptions broken out by category. Since there are a large number of variables for binary indicators, we need to do feature engineering to reduce the number of features. Other

integer variables include cms_disabled_ind (Disability Indicator), cms_hospice_ind (Hospice Indicator), cms_low_income_ind (Low Income Subsidy Indicator from CMS), est_age (Member age calculated using est_bday, relative to score/index date), and so on.

2.2 Data Preprocessing

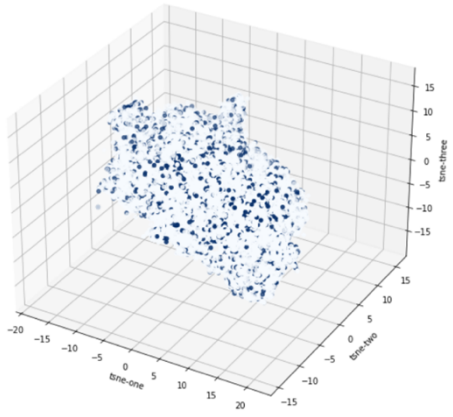
Firstly, we cleaned both datasets by eliminating abnormal symbols and converting value types.
Then, we set up rules of missing values imputation according to the distribution and value type of each variable in both training dataset and holdout data.

- If the mean values are very close in both datasets, we imputed missing values with the mean value, else we impute with the median.
- If the number of unique values of one variable is less than 30, we imputed the missing values by the most frequent value.
- For the object variables, we imputed the missing values with the most frequent value.

index	column type	null values	null values(%)
cms_low_income_ind	int64	0	0
cms_ma_risk_score_nbr	float64	3772	5.421721382
cms_partd_ra_factor_amt	float64	3814	5.482090496
cms_ra_factor_type_cd	object	4224	6.071408038
cms_risk_adj_payment_rate_a_amt	float64	3748	5.387224746
cms_risk_adj_payment_rate_b_amt	float64	3750	5.390099465
cms_risk_adjustment_factor_a_amt	float64	3751	5.391536825
cms_rx_risk_score_nbr	float64	3749	5.388662105
cms_tot_ma_payment_amt	float64	3750	5.390099465
cms_tot_partd_payment_amt	float64	3750	5.390099465

3. Feature Engineering

We applied PCA+TSNE methods to reduce the dimension of raw data and observe the distribution of two different categories. Unfortunately, after reducing the dimension to 50 degrees and applying TSNE to visualize the data, we still cannot tear apart the data in different categories. Positive data and negative data are still mingling with each other, which means current features are not able to detect the label. Consequently, we are implementing feature transformation and feature selection methods to generate features which can lock on the positive category and reduce the dimension of the dataset.



3.1 Feature Transformation

We applied various methods to generate features.

- Polynomial expansion: add polynomial terms
- Correlation detection: find correlated variables and delete the one with less information
- Importance detection: combine important features with other features
- Feature combination: combine similar features
- Domain knowledge: combine features with intuitive thoughts

3.2 Feature Selection

After having all features generated, we applied different feature sets by using different methods, and then we tested the performance of each feature set and selected the best set.

Filter Method	Variance	Remove features that show the same value for the majority/all of the observations (constant/quasi-constant features)
	Correlation	Remove features that are highly correlated with each other
	Chi-Square	Compute chi-squared stats between each non-negative feature and class
	Information Value (IV)	IV = Σ(Proportion of Good Outcomes - Proportion of Bad Outcomes) * WOE
Embedded Method	Lasso (L1)	Regularization consists in adding a penalty to the different parameters of the machine learning model to reduce the freedom of the model.
	Random Forest Importance	
	Gradient Boosted Trees Importance	
Feature Shuffling	Random Shuffling	Shuffling the values of each feature, one at the time, and measure how much the permutation decreases the roc_auc of the machine learning model.
Hybrid Method	Recursive Feature Elimination	Rank the features according to their importance Remove/add one feature -the least important- and build a machine learning

		algorithm utilizing the remaining features Calculate a performance metric of roc-auc
	Recursive Feature Addition	
Heuristic Rules	Domain knowledge	

Take the Heuristic method as an example:

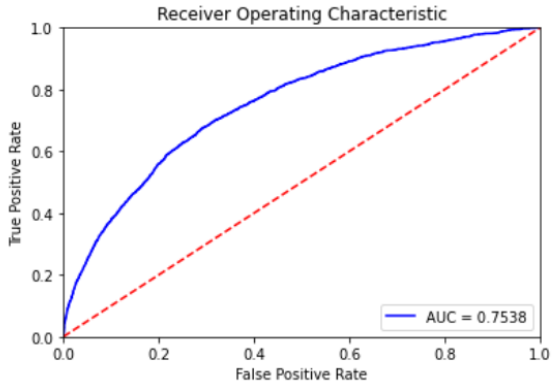
1. We summarize all the variables and delete highly correlated variables which contain less information.
2. Since some features use two variables to store its values, for example, betos_d1c_ind and betos_d1c_pmpm_ct. However, they are highly correlated, so we deleted the indicator.
3. We delete zip_cd and cnty_cd because there should be too many dummies if we keep them. And zipcode and country will not provide a lot of information for our prediction.
4. After these three steps, we reduce the number of variables from 824 to 488.

4. Modelling

We chose various models, including Logistic Regression, SVM, Random Forest, XGboost, MLP, Sequential, to do classification, and we use the best feature set to evaluate the performance of each model. We chose the result of Logistic Regression as the baseline – AUC 0.66.

Our problem is a binary classification problem. The top performing model was a XGBoost gradient boosting tree classifier with hyperparameters: base_score = 0.5, booster = "gbtree", colsample_bylevel = 1, n_estimators=40000 max_depth=4, learning_rate=0.001, subsample=0.4, colsample_bytree=0.6, min_child_weight = 5, missing=-1. A gradient boosted tree is a supervised learning technique, which produces a prediction model in the form of an ensemble of weaker prediction models (i.e. decision trees). According to the validation results, the XGboost can achieve the AUC of 0.7358 on test dataset. Here is the AUC curve graph.

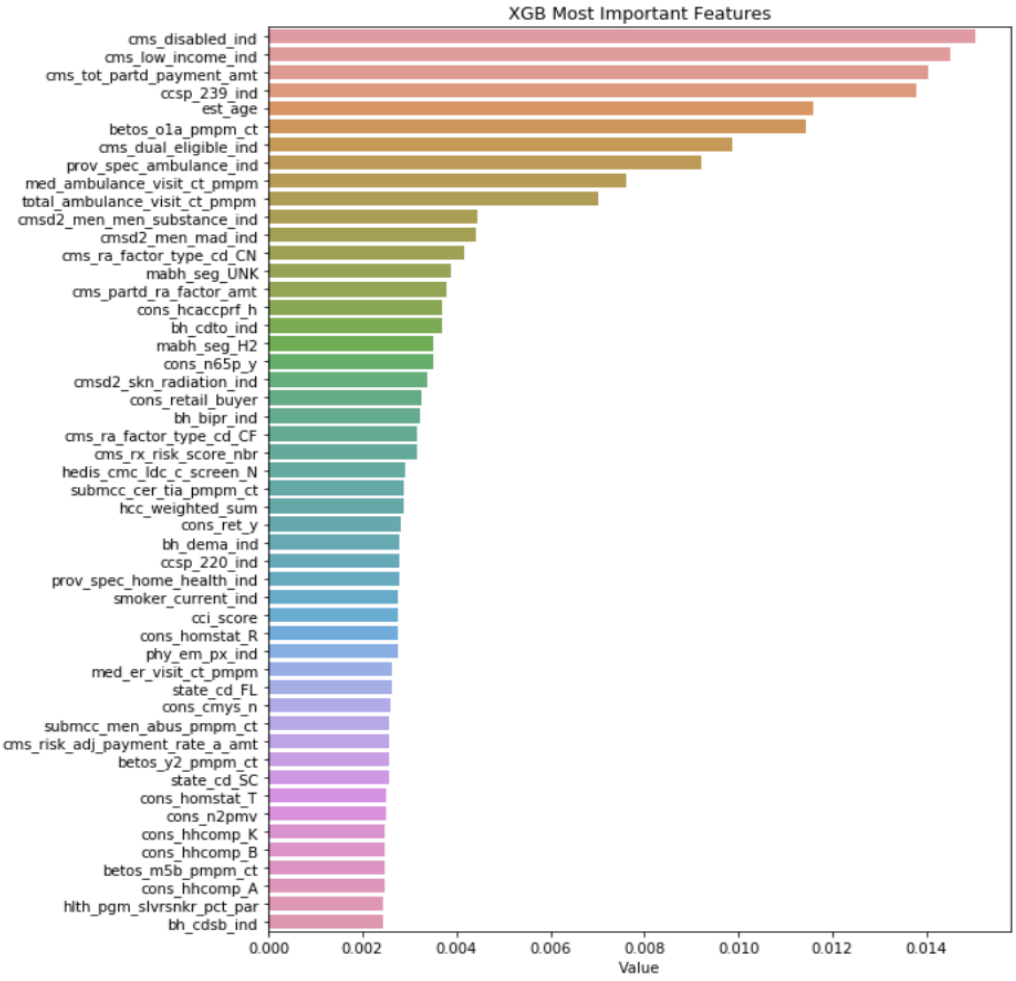
```
ROC_AUC Score:
0.75380205973414
```



5. Key Performance Indicator Analysis:

5.1. Overall Feature Importance

Below is the importance plot of top 50 features.



Below is the importance summary of the top 10 features.

Rank	Importance Value	Correlation with Target	Feature	Description
1	0.01502447	0.1586	cms_disabled_ind	Disability Indicator
2	0.0145182	0.1724	cms_low_income_ind	Low Income Subsidy Indicator from CMS
3	0.01403961	0.1981	cms_tot_partd_payment_amt	Total PartD Payment Amount
4	0.01378314	0.1770	ccsp_239_ind	Binary indicator for CCS code - Superficial injury, contusion

5	0.01160058	-0.1821	est_age	Member age calculated using est_bday, relative to score/index date
6	0.01144293	0.0940	betos_o1a_pmpm_ct	Per Member Per Month Count of Logical Claims for each of the BETOS codes
7	0.00985391	0.1572	cms_dual_eligible_ind	Dual Eligibility Indicator - eligible for both Medicare and Medicaid
8	0.00922127	0.1646	prov_spec_ambulance_ind	Binary indicator for a select group of categories using std_hipaa_prov_spec_cd
9	0.00760557	0.0987	med_ambulance_visit_ct_pmpm	Per Member Per Month Visits for non-BH related claims, broken out by utilization category
10	0.00700174	0.1190	total_ambulance_visit_ct_pmpm	Combined Med & BH values from above, broken out by utilization category

5.2. Grouping features in 4 main categories

The top 10 most important features fall into 4 categories, which are CMS Features, Demographic/Consumer Data, Medical Claims Features, Other Features.

6.

CMS Features	(1) cms_disabled_ind (2) cms_low_income_ind (7) cms_dual_eligible_ind
Demographic/Consumer Data	(3) cms_tot_partd_payment_amt (5) est_age
Medical Claims Features	(4) ccsp_239_ind (6) betos_o1a_pmpm_ct
Other Features	(8) prov_spec_ambulance_ind (9) med_ambulance_visit_ct_pmpm (10) total_ambulance_visit_ct_pmpm

Observations and Recommendations

Observation 1: cms_disabled_ind and cms_low_income_ind are the top 2 most important indicators influencing the transportation issue. Both indicators have a correlation of over 0.15 to the target variable and a 1% importance value in the model. est_age is the fifth most important feature, which has a negative impact on the target variable.

Recommendation 1 -- Offering Scheduled Bus

The result is consistent with our intuition: people with disability and people in lower income range are more likely to have transportation issues. This implies that physical disability and financial disability are the key factors that lead to transportation difficulties. There is a negative correlation between low income

indicator and estimated age, which shows that younger people are more likely to be poor and to face transportation difficulty.

After careful consideration, we believe offering special service targeting this group can be feasible: we can schedule a bus charging minimum cost every week travelling around specific neighborhood where people are more likely to have transportation issues and taking them to the hospital.

Observation 2: ccsp_239_ind indicates Superficial injury and contusion, which is the fourth most important feature to influence transportation issue. It has a positive impact on the transportation issue.

Recommendation 2 – "Small Ambulance" for Superficial Injury and Contusion

Normally patients with Superficial injury and contusio will seek for help in emergency care or clinic service. They might have difficulty transporting because in those situations they are in a dilemma: costly ambulance and difficulty transporting to the hospital. They probably did not and were not willing to ask help from Taxi or Uber. It will be a possible solution to build up the pickup system for superficially injured patients. Less equipped "ambulance" with less priority and lower pricing could be the best choice for patients falling into this category.

Observation 3: As we noticed, ambulance issues account for several most important features determining the transportation issue, including betos_o1a_pmpm_ct, prov_spec_ambulance_ind, med_ambulance and total_ambulance. They all positively impact the target variable.

Recommendation 3 – Ambulance Issue

Under most situations, people who visited the hospital by ambulance were seeking for emergency service. Thinking practically, their claim on transportation issue might result from slow or relatively high cost ambulance service. This is no direct or efficient solution for this problem coming right out of hand. However, with the predictive model on transportation issue, it could be a possible solution to set up a new system and pricing strategy for people who might need multiple ambulance services.

A certain group of people with low credit scores might use the ambulance system inappropriately to seek faster service from the emergency without paying the bill. This kind of behavior might negatively impact the normal operation of the hospital. Recognizing these people and coming up with better solutions could benefit both Humana and the clients.

Observation 4: cms_tot_partd_payment_amt is the third most important feature. It has a positive contribution to the transportation issue.

Recommendation 4 – Transportation coupon and subsidies for Part D members

The Medicare Part D program began providing prescription drug coverage to Medicare beneficiaries in 2006. In January 2020, the Centers for Medicare & Medicaid Services (CMS) Center for Medicare and Medicaid Innovation (Innovation Center) began the Part D Payment Modernization Model to test the impact of a revised Part D program design and incentive alignment on overall Part D prescription drug spending and beneficiary out-of-pocket costs. The Model aims to reduce Medicare expenditures while preserving or enhancing quality of care for beneficiaries.

People with chronical diseases tend to spend more money for medical treatment. From the result, we see that people with more medical expenses will face more transportation challenges. This is

probably because chronic disease care requires frequent clinician visits, medication access, and changes to treatment plans in order to provide evidence-based care.

Therefore, from the Part D payment database records, Humana can easily identify those with large amount of medical expenses. These people have a high probability in facing transportation challenges. Humana may collaborate with Uber, Lyft, or public transportation departments, such as providing some coupons to this group of people to help them conquer the transportation issues.

Also, we find that there is a very high correlation (0.8248) between low income indicator and Part D payment amount. We made the assumption that large amount of medical expenses leads to poverty. Therefore, helping these people by providing transportation subsidies is very important.

7. Overall Takeaways

Following are our overall takeaways:

- We developed machine learning predictive model and detected features significantly indicating transportation issues as shown in part 5.2.
- We recommend Humana offering scheduled bus for people who have finance difficulty or disabilities to transport to the hospital.
- We recommend Humana offering inferior ambulance service with low cost and less priority for patients with superficial injuries or contusions.
- We recommend Humana detecting patients who do not use ambulance system properly (people who try to get access to emergency service faster in an inappropriate way) and set up new system for these patients.
- We recommend Humana providing patients who have chronic diseases and frequent medical needs with transportation services such as collaboration with public transportations.

8. Appendix