Import Identifiers and Prediction about Transportation Challenges with Logistics Regression

Team Procrastination

"...During this pandemic period, Humana is doing everything she can to assist and protect health as humans care for each other..."

Introduction

With the rapid pace of social development, we humans have enjoyed an easier way of living and commuting. Meanwhile, we are exposed to a growing number of health-threatening substances and sitee, which makes it harder for the medicare industry to evaluate health and its cost. Fortunately, the introduction of Humana's integrated value-based health ecosystem based on Social Determinants of Health brings a promising prospect to this industry. This report is intended to find the interconnection between our socio-economic and community environments and lifestyle behaviors, based on which employees, patients, clients are better medically assisted.

Among all those challenges to health, transportation challenge stands out to be the most common one since it involves daily life of a person's commuting, schooling, and living. The pain point lies in the aspect of prediction and optimal choices of model fitting. With a large amount of robust data provided by Humana, this report will accurately identify medicare members who are most likely to experience Transportation Challenges and propose viable solutions.

The subsequent sections will mainly cover summary of findings, exploratory data analysis, model development and assessment, and business recommendations.

Summary

From 893 features, we utilize sklearn feature selection to select out 130 most important features and fit them with a sklearn logistic regression model with alternating seeds to ensure not overfitting the data. Our selected features cover 21 categories, inside which different minor categories are also distinguished. As is shown in Table 1.1, these categories are mainly evaluating the probability from the aspects of financial situations, health conditions, and demographic groups.

Generally speaking, an aged female member with disability or cardinal/lung illness are most likely to experience transportation challenges. The varying medicare cost (*BETO*) of employees also leads to different behaviors of those members and different likelihood of experiencing transportation challenges. In addition, *credit* issues are also significant but the actual threshold for the danger zone of credit is not firmly corroborated for lack of patterns. Last but not least, the interconnection between participation in *health programs* and *medical/drug prescriptions* suggests that providing or encouraging members to participate in health programs is a good practice for reducing risk of transportation challenges.

Variable Name	Description
sex_cd	Member gender
est_age	Member age calculated using est_bday, relative to score/index date
ccsp*	Clinical Classification Software
cms*	Centers for Medicare and Medicaid Services
cmsd2*	CMS Level 2 diagnosis categories
cons*	Amerilink Consumer Date from the KBM Group
credit*	% Balance to a financial variable
hedis*	Healthcare Effectiveness Data and Information Set
hth*	Health program
lab*	Abnormal Lab Results broken out
pdc*	Percent Days Covered - a measure of medication adherence
phy*	Physican E & M
prov*	prov_line_cd

rev*	Revenue code CMS categories
submcc*	subcategory of major clinical category
betos*	Barenson-Eggers Typeof Service Codes
rx*	Drugs
lang_spoken_cd	Preferred Language for Member
cci_score	Charlson Comorbidity Index value, sum of clinical and age components
dcsi_score	Final Diabetes Cormorbidity Severity Index score
fci_score	Functional Comorbidity Index value calculated from 18 components
zip_code*	Zip code

*Note: * means variables beginning with the string before it*

Table 1.1 Feature categories and detailed descriptions

Detailed recommendations are given in the Recommendations Section. Also there are certain features that do not contain any positive label so they are simply excluded. However, as the number of members and methods of transportation change, the recommendations are subject to change.

Exploratory Analysis

Missing values

The data we have has missing values. More importantly, some variables have too many missing values to do further analysis. Therefore we have to look into those variables and handle them in a meaningful imputation method.

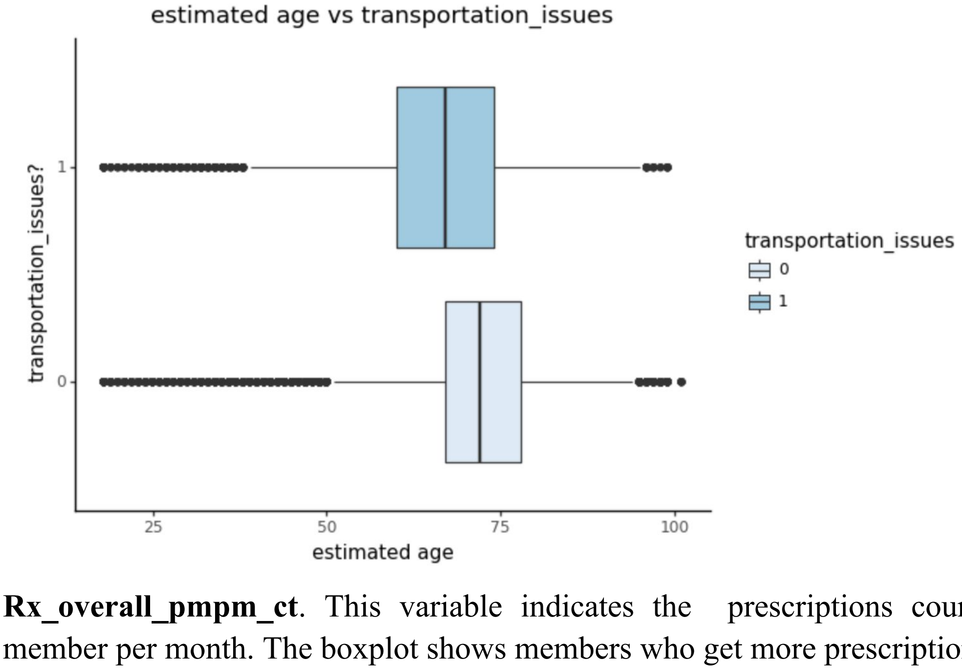
Categorical variables

22 of the 826 variables are categorical and contain string values. To fit our model using the sklearn python library, we need to encode these variables to numerical values. Our further observation of these variables also indicates none of them are ordinal and intuitively One-Hot encoding becomes the best encoding strategy.

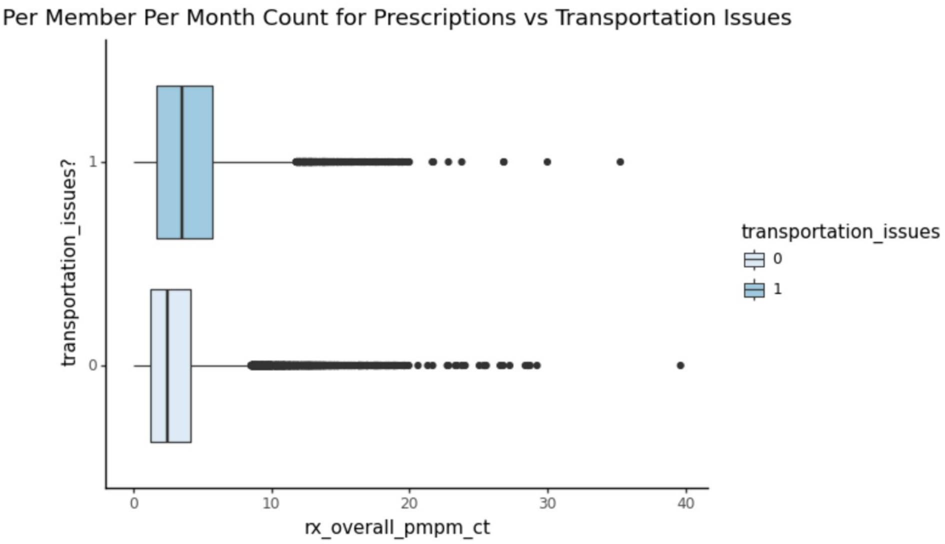
Important variables

As we started our exploratory analysis, we found a substantial amount of variables worth further investigating. In this section, we present four variables that are most interesting: *est_age*, *rx_overall_pmpm_ct*, *sex_cd*, and *lang_spoken_cd*.

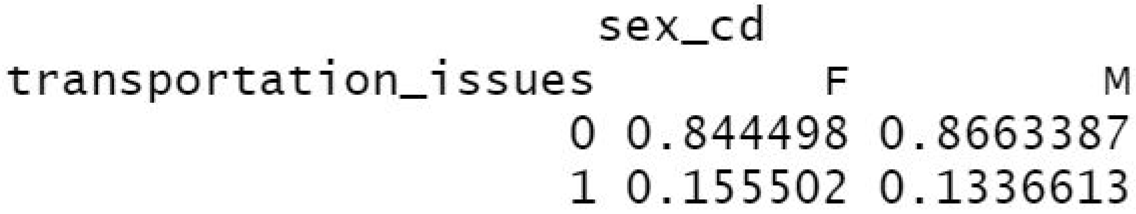
- **Est_age.** The boxplot of estimated age vs transportation_issues shows a quite counterintuitive result. It is intuitive to think older people will be more likely to have transportation issues. However, the data shows the opposite. Both the median and the distribution are suggesting that people who have transportation issues are younger than people who do not.



- **Rx_overall_pmpm_ct**. This variable indicates the prescriptions count for each member per month. The boxplot shows members who get more prescriptions are more likely to have transportation issues.

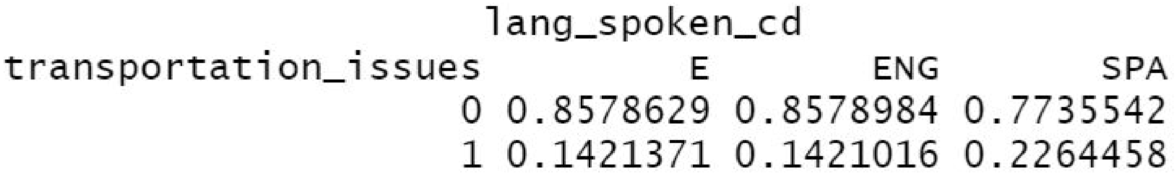


- **Sex_cd.** Our data suggest female members are 2.2 % more likely to have transportation issues than male members.



```
                          sex_cd
transportation_issues         F         M
                    0 0.844498 0.8663387
                    1 0.155502 0.1336613
```

- **Lang_spoken_cd**. Surprisingly, members whose preferred language is Spanish are 8% more likely to have transportation issues than members whose preferred language is English or other.



```
                     lang_spoken_cd
transportation_issues         E       ENG       SPA
                    0 0.8578629 0.8578984 0.7735542
                    1 0.1421371 0.1421016 0.2264458
```

To conclude, by visual inspection, these four variables are of great importance. They will play important roles in predicting transportation issues. However, these findings are based on plots. Their statistical significance will be determined in the more statistically rigorous model selection and feature engineering section.

Model

Imputation

Imputation is the process of filling missing values in a dataset based on the statistical measurements such as mean, mode, or median of other values. Through initial spection of the dataset, we found lots of issues such as inconsistent data types as well as inconsistent imputation of missing values, which made it difficult to clean and interpret data. The accompanied Data Documentation indicated that there were some some 'missing values' are imputed in an undesired fashion, and the table below summarizes things that we needed to be mindful of prior to performing imputation:

Issues	Sample Predictors	Strategy
'*' is used in certain predictors to represent synthetic artifact	● *cms_ra_factor_type_cd* ● *cons_cmys*	Replace with np.nan to classify as missing values
'other' is used in certain predictors to represent synthetic artifact	● *zip_cd* ● *cnty_cd*	Replace with np.nan to classify as missing values
1.1 is used to indicate an 'not applicable' scenario	● All PDC features	Convert features to binary, where 1.1 is encoded as 0, all other values are encoded encoded as 1
'UNK' is used to certain predictors to indicate missing values	● *mabh_seg*	Replace with np.nan to classify as missing values
'U' is used in certain predictors to indicate unknown	● *cons_hhcomp* ● *cons_homstat*	Replace with np.nan to classify as missing values
'0' is used in certain predictor to indicate unknown	● *cons_cmys*	Replace with np.nan to classify as missing values

Using the table above, our team first made corresponding value conversions to ensure that all missing values are appropriately represented as 'nan' so that imputation could be carried out through the entire dataset.

To perform imputation, our team believed it was not sufficient to take column median or mean, since the amount of missing values in certain groups occupy close to ⅔ of the observations. Therefore, a more sophisticated method was required. More specifically, we chose to impute a column with missing values with a related column that does not contain missing values. For example, the predictor '*zip_cd*' encoded geographical information and contained over 45000 missing values. However, it conveyed similar information is *rucc_category*, which stands for Rural Urban Continuum Code and could be interpreted as the population size of a location. And we made an assumption that these two predictors are related. Therefore, we could group the dataset by *rucc_category*, and impute the missing zip codes within each rucc category with the most frequently recorded zip code for that particular category. This idea was used as the basis of our imputation strategy, and is concluded in the table below:

Source Predictor	Target Predictor
rucc_category	● *zip_cd*
zip_cd	● *cons_n2mob, cons_n2pbl, cons_cmys, etc.*
hcc_weighted_sum	● All cms columns
cci_score	● *cons_hcaccprf_h, cons_hcaccprf_p, etc.*
dcsi_score	● All hedis columns
est_age	● *cons_n2029_y, cons_n65p_y*

One thing worth noting is that there were 840 unique zip codes, and only 9 rucc categories. It did not make sense to group 840 unique values by just 9 groups. So we only downsampled *zip_cd* by using only the first 2 digits, which retained geographical information but significantly reduced categories. This process also made encoding of *zip_cd* less burdensome. During the imputation, we also dropped some variables that contained redundant information, this included:

- *cnty_col, state_col, person_id_syn, srs_platform_cc, etc.*

Encoding

Encoding refers to the process of converting each level of a categorical variable to discrete numeric values, which is a required step in order for statistical packages such as scikit-learn to parse the data. Categorical data can generally be divided into ordinal categorical data and normal categorical data, and should be treated differently when performing encoding.
More concretely, due to the intrinsic ordering presented in ordinal categorical data, these variables could be encoded to integers ranging from baseline level such as 0 to the maximum levels of these predictor variables. For nominal variables, one-hot-encoding is typically performed. This process would also produce extra features in the dataset, where the number of additional features equal to the number of unique values of the encoded variable.
We summarized the categorical features in the table below:

Ordinal	● *cons_cmys, mabh_seg, cons_homstat, rucc_category*
Nominal	● *sex_cd, lang_spoken_cd, cms_ra_factor_type_cd, cons_hhcomp, hedis_dia_eye, hedis_dia_hba1c_ge9, hedis_dia_hba1c_test, hedis_dia_ldc_c_control, hedis_dia_ldc_c_screen, hedis_dia_ma_nephr, zip_cd*

Predictor *mabh_seg* represented medicare segmentation and contained 15 discrete levels. We however collapsed this predictor into 2 levels - 'Healthy' and 'Chronic'. Predictors *cons_cmys*, which represented education level and *rucc_category* were inherently ordinal. We also decided to treat cons_homstat (home ownership status) as an ordinal variable, where 'renter' was encoded as the baseline, and 'homeowner' was considered the highest level.

The encoding of nominal variables were straightforward. We ensured that there were no missing values and we used the *category_encoders* package for one-hot encoding.

Feature Selection

During EDA, we noticed that certain groups contained both the binary indicator and the percentage value of the same feature. For example, *betos_1dc* from the betos group has a binary indicator and a per member per month count. We thought such information is redundant and our predictive model should not require both. Therefore, the initial step of feature selection was to drop numeric columns that contained such duplicates. This generally included predictors with the following suffix:

- *_pmpm_ct*
- *_days_pmpm*
- *_amt*

We also noticed that certain binary predictors contained the same value across all observations, which could result from the imputation we performed earlier. So we dropped these predictors.

With the remaining predictors, we decided to go through them individually and inspect their relationship with the response variable for any obvious patterns. Specifically, for discrete predictors, we grouped them by *trainsportation_issue* and check for data distribution across different levels of the predictor. An example of this is depicted in the table below:

Predictors	No Transportation Issue	Has Transportation Issue
'betos_o1e_ind'	6.694451175187788	4.910484668644906
'betos_t2a_ind'	6.434704830053668	5.058123817247905

'betos_t1a_ind'	4.526752767527675	6.275528902215559
'betos_m2c_ind'	5.938983488132095	3.4534606205250595
betos_m5b_ind	6.178183816458304	2.819506726457399
...		

The table above uses betos columns as an example. No Transportation Issue referred to the ratio of positive betos cases over negative betos cases where observations claimed no transportation issues, and likewise for the Has Transportation Issue column. We then look for predictors where the difference between these 2 columns are greater than a threshold such as 2.5, indicating a pretty big change between the two response groups. Or predictors where one value is greater than 1 and the other is less than 1, indicating a completely different trend between the two response groups. In the example above, all predictors except for *betos_m5b_ind* missed our selection criteria, and were therefore dropped. This same idea was carried out for all binary predictors.

For numeric predictors such as those associated with credit, we used a similar strategy, except that the mean value of each response group was used instead of the ratio used for binary predictors. We then chose predictors where the differences between the response groups were greater than 10%. One example of such is shown below:

Predictors	No Transportation Issue	Has Transportation Issue
credit_bal_agencyfirstmtg_new	5806.567716196336	5189.736019409852

This selection process significantly reduced our feature size to just around 200. We then used the SelectKBest class from scikit-learn and chose f_classif (ANOVA f-test) to further narrow down our predictors. More specifically, we test an array of different feature sizes, ranging from 50 to 150 with an increment level of 10, and AUC as the evaluation matrix as well as cross validation with 10 different folds to determine the optimal feature size. This is detailed in the table below (using Logistic Regression):

Feature Size	Average AUC (10 folds)
50	0.7338490179791616
60	0.7354861993808879
70	0.7379715204394868
80	0.7378735763519708
90	0.7376050981355246

100	0.7377763437675945
110	0.7387443116568815
120	0.7389109079703755
130	0.7391143641122839
140	0.7389146748659883
150	0.7387529433620007

As a result, we selected the top 130 features as returned by the previous step.

Model Selection

To select the best model, we considered three classifiers: Logistic Regression, Random Forest, and XGBoost. We first ran each model on our training set with cross validation using 10 folds. The performance of each model and parameters are listed below:

Mode	Average AUC (10 folds)
Logistic Regression	0.739
Random Forest	0.720
XGBoost (n_estimators = 1000)	0.736

Compared to the AUC scores the other two models yield, Logistic Regression from scikit-learn had the best performance and was used as our final model. We also tried to tweak the parameters to maximize our model performance without over fitting. Below is a summary of our final model:

Model	LogisticRegression
penalty	L1
solver	liblinear
max_iter	150

Output Generation

To generate the final submission csv file, we fit our model on the hold out file and used predict_proba function to generate a score for each ID. Next, we dropped all columns except

ID and created a column “RANK” and assigned ranks by the scores. For example, if we have three employees: A: 0.9, B:0.9, C:0.8, the rank they get are: A:1, B:1, C:2.

Recommendation

Most important identifiers include *sex, age, credit, submcc, spoken language, zip_code, betos code*, etc. It is important to mark down members who are female or live distantly from the workplace or in *rural* areas. Also, members who are more than 50 years old or have cardinal/lung illness are more likely to experience transportations challenges.

In addition, the *credit* or financial situations are also an important factor as it represents a member's affordability of safe vehicles or clean and friendly living environment. However, there is indeed little that Humana can do to help these members in credit risk other than vouching for them, which is not a realistic option.

Last but not least, the language a member spoken may also affect the likelihood of transportation challenges. This is probably because of the skewed number of English speakers and Spanish speakers. But according to Chen, *language spoken* does affect humans' social behaviors and economical habits, which can be better studied with a larger dataset.

Therefore, given the recommended identifiers above, Humana can take the following further actions or give suggestions to its members.

Suggest female members who do not drive to work and live in distant/rural areas arranging pickups or asking for Uber/Lift allowance from companies.

According to Gautam and et. al., sexual harassment in public transportation is much more frequent to female passengers who travel along in distant areas and it can incur great mental stress to the victims. Arranging pickups or issuing Uber/Lift allowance can help protect the interests of female members and reduce the risk of mental illness.

Suggest aged members asking for loose on the on-time attendance requirement from employers.

According to Bett and et. al., public transportation failed to detect and prepare the fast transit of passengers with cardinal illnesses. Among those passengers, aged patients with heart attack are the most vulnerable groups. Since a rush to work is much likely to trigger minor ventricular fibrillation or heart attack, a loose on the attendance requirement on these employees will be a good choice.

For Humana's own employees, Humana can move working hours several minutes later or earlier than companies near Humana so that Humana employees do not rush to work at the same time employees from other companies do.

According to Abdallah in the book *Sustainable Mass Transit: Challenges and Opportunities in Urban Public Transportation*, there are numerous minor factors in daily public transportation that can affect the health of commuters. Among them, the most important one is the risk of exposure to contagious virus and flu especially during this pandemic time period. And according to Sinha in *Journal of Transportation Engineering*, it normally takes 10 to 30 minutes for a regional rail to hold and finish transiting the group of passengers that arrive at the railway platform at the same time. Therefore, shifting work hours and making it stagger with other companies' can reduce the risk of virus exposure and unexpected infection.

Pay attention to members who go to work by public transportation or do not own a private vehicle at all.

According to Chaturvedi and Kim in *Long term energy and emission implications of a global shift to electricity-based public rail transportation system*, emissions from bus and agricultural vehicles contain larger portions of harming compounds than that from compact vehicles. Employees who drive to work in a sealed space are less likely to inhale those poisonous compounds and the only thing that may affect their health is mental torment and deduced wages from traffic jams.

Conclusion

Generally speaking, the recommendations given above still need to be considered together with legal issues and regulations. Through careful imputation, cross validation, and randomForest, the auc of our final model is approximately 0.739 and there is little issue of overfitting or underfitting. The most important identifiers include sex, age, credit, submcc, spoken language, zip_code, betos code, etc. Afterwards, Humana can adopt those recommendations and gain a better overview about the likelihood of transportation challenges that its members encounter.

Reference

Abdallah, T. (2017). *Sustainable Mass Transit: Challenges and Opportunities in Urban Public Transportation*. Elsevier.

Bett, J. H. N., Tonkin, A. M., Thompson, P. L., & Aroney, C. N. (2005). Failure of current public educational campaigns to impact on the initial response of patients with possible heart attack. Internal medicine journal, 35(5), 279-282.

Chaturvedi, V., & Kim, S. H. (2015). Long term energy and emission implications of a global shift to electricity-based public rail transportation system. *Energy Policy*, *81*, 17.

Chen, M. K. (2013). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. American Economic Review, 103(2), 690-731.

Gautam, N., Sapakota, N., Shrestha, S., & Regmi, D. (2019). Sexual harassment in public transportation among female student in Kathmandu valley. Risk management and healthcare policy, 12, 105.6-185.

Sinha, K. C. (2003). Sustainability and urban public transportation. *Journal of Transportation Engineering*, *129*(4), 331-341.