Humana Mays Case Competition 2020

By,

Soumya Patra

Nikhil Viswanath

Table of Content

Introduction

"*The field of public health refers to the conditions that are not medical but that can produce or undermine health as the "social determinants of health." These are the socioeconomic, environmental, and behavioral factors that research over many decades has shown to be strong influences on health."* **— Elizabeth H. Bradley, The American Health Care Paradox: Why Spending More Is Getting Us Less**

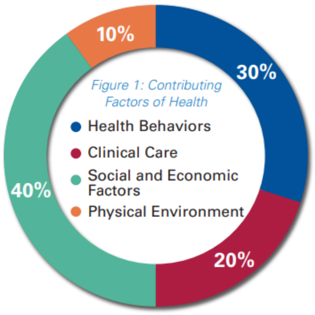

Figure 1

Healthy well beings of people are linked to the social and economic conditions in which they leave. Prior research by Health Research & Education Trust, 2017 state that 20% of a person's health can be contributed to the medical care whereas 40% of it are attributed to healthy food, housing status, educational attainment and access to transportation.

Individuals struggling with food insecurity, housing instability, limited access to transportation or other barriers may experience poor health outcomes, increased health care utilization and increased health care costs. Addressing these determinants of health, commonly referred to as social determinants of health will have a significant positive impact on people's health, including longer life expectancy, healthier behaviors and better overall health.

Transportation is an economic and social factor that shapes people's daily lives and thus a social determinant of health. Transportation is interrelated with other social determinants of health such as poverty, social isolation, access to education and racial discrimination. Barriers to transportation greatly affect the quality of people's lives. These statistics highlight the scope of the problem:

- 3.6 million people in the U.S. do not obtain medical care due to transportation barriers
- Regardless of insurance status, 4 percent of children (approximately 3 million) in the U.S. miss a health care appointment each year due to unavailable transportation; this includes 9 percent of children in families with incomes of less than $50,000
- Transportation is the third most cited barrier to accessing health services for older adults

Transportation challenges affect urban and rural communities. Overall, individuals who are older, less educated, female, minority, or low income—or have a combination of these characteristics—are affected more by transportation barriers.

Children, older adults and veterans are especially vulnerable to transportation barriers due to social isolation, comorbidities, and greater need for frequent clinician visits. Transportation issues affect people at varying levels depending on how different challenges overlap.

For example, a low-income person struggling with travel may have an increased burden if he or she experiences a temporary physical disability. Limited health literacy, cognitive impairment, fragmentation of health history, access to health insurance, poverty or food insecurity can intersect at any period and affect individuals and communities.

Transportation issues include lack of vehicle access, long distances and lengthy travel times to reach needed services, transportation costs, inadequate infrastructure and adverse policies that affect travel. Transportation barrier are due to the following reasons.

Transportation Infrastructure	Transportation Cost	Vehicle Access	Distance and Time Burden	Policy	Impact on Health Care Access and Health
• Limited availability and routes • Overcrowding on public transportation • Roads and transport stations in disrepair • Safety issues	• High cost of fares • Personal vehicle expenses such as insurance • Credit card or bank account requirements	• Lack of a personal vehicle • Lack of access to a vehicle through friends or family	• Long travel distances and lengthy wait times • Erroneous or inconvenient time schedules	• Budget cuts resulting in bus and train shortages, routes removed, and strikes • Driver's license barriers • Lack of adequate transit in underserved areas	• Missed doctor and clinic appointments • Limited pharmacy access and decreased prescription fills • Economic burden for patients and the health care system

Table 1: Reasons for transportation barrier

Impact on Health Care Access and Health
• Missed doctor and clinic appointments • Limited pharmacy access and decreased prescription fills • Economic burden for patients and the health care system

Table 2: Impact of transportation barrier on healthcare

Methods

With a primary aim to perform both descriptive and predictive analytics on the provided data, the study will follow the below approach for analysis:

Data Preparation
1. Correction of labels where there are miscoding from the source data. Example: for the column language spoken label marked as ‘E’ is recoded as’ENG’. 2. Setting the data type for the variables as interval variable or categorical variable 3. Converting labels for variables where synthetic data was used to missing values. Example: for the column with label as 1.1 was converted to Null value
Feature Engineering
1. Imputation of missing values to mean for interval variable and mode for categorical variable 2. Variable transformation for variables which do not have normal distribution or have outliers
Feature Selection
1. Use Light Gradient Boost to select variable based on feature importance using imbalanced data 2. Use Gradient Boost to select variable based on feature importance using balanced data, balanced data created based on up sampling using synthetic data for minority class (SMOTE) 3. Removal of indicator variable for which we have continuous variables 4. Removal of categorical variable with more than 80% in one factor 5. Removal of interval variable with high skewness 6. Removal of variables with more than 50% missing values
Modelling
1. Data was split into train and test using stratified sampling with a ratio of 70/30 for train and test 2. Based on features selected, following models were performed like Random Forest, Gradient Boost, Light GBM, XGBoost, SVM and Auto Neural Network 3. The modelling was based on predicting transportation issues and the best model was selected based on the highest AUC on the test data 4. To avoid overfitting cross validation was performed on 5 folds

Scoring/ Generalization
1. Best model selected based on test AUC was used score the holdout data
2. The predicted probability was used to rank the patients in the data

Table 3: Steps of analysis

Analysis and Findings

Descriptive Analysis

The dataset provided was a combination of various demographic, medical, census and diagnostic information where the granularity is defined at an individual member level. There were two different datasets provided where one was a training dataset which had labeled information of members who were facing transportation issues and another dataset which was a holdout sample used to score the predictive model. The training set consisted of ~69,000 records with 14.6% of the records labeled as having transportation issues.

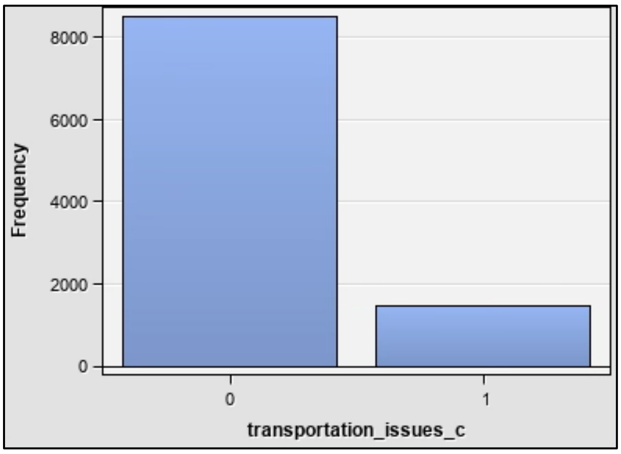

Figure 2

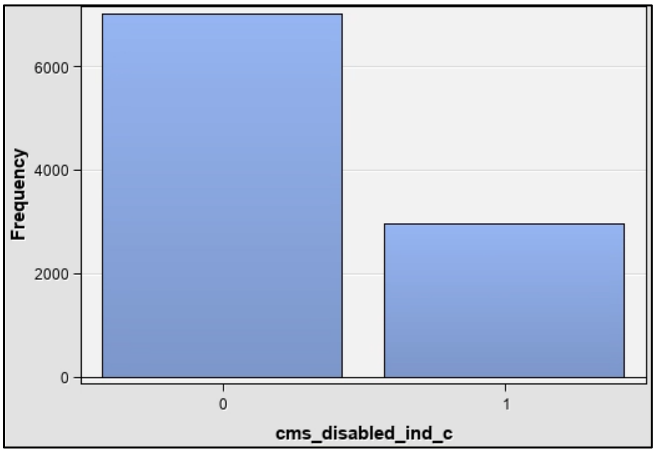

Figure 3

This clearly indicated that there was an imbalance in the dataset and presented a challenge in applying predictive modelling techniques without compensating for sampling measures. The disability indicator (cms_disabled_ind) was another important metric available in the data and revealed that ~21,000 records contained members with some form of disability. It is important to note that this was a survey conducted by Humana across its wide range of customers and therefore there were a lot of features included in the dataset. Around 825 features were included in the dataset like information about a member's credit history, medical history, amount of prescription drugs consumed, census survey questions and various medical screening test history over the last few months. Age was another critical factor in the dataset and from the below distribution we can see that the average age for most members in the datasets was between 60-75 years. Older people are also more prone to have transportation issues making it a crucial variable in our analysis.

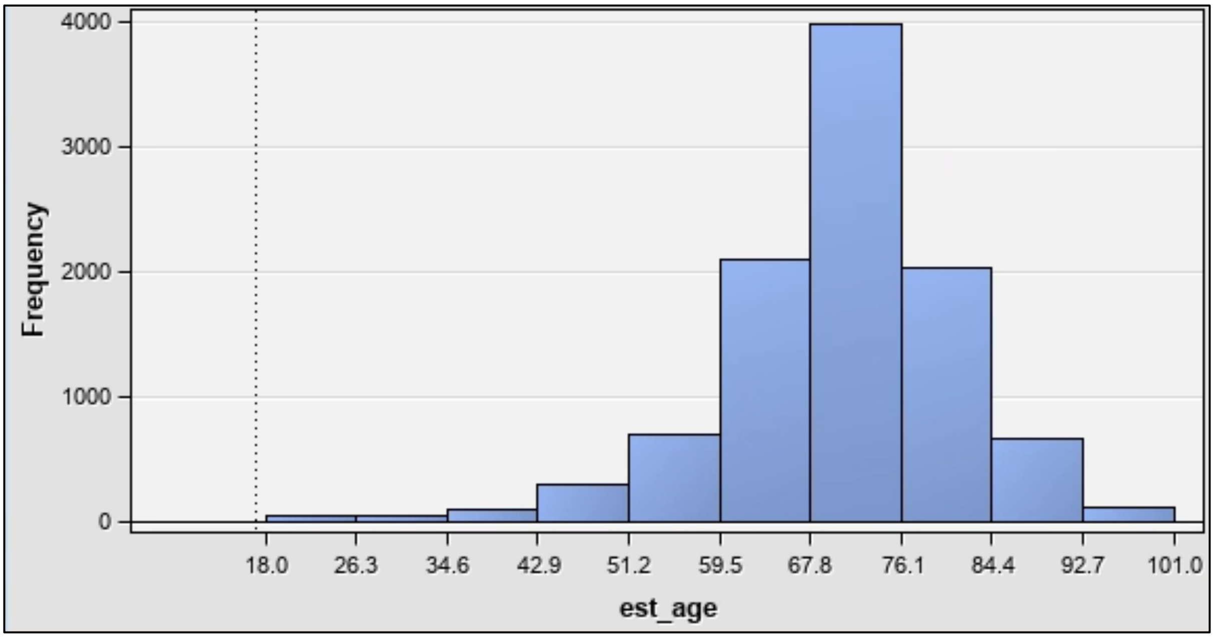

Figure 4

Our biggest challenge was to reduce the dimensionality in the data and ensure we are working with a smaller dataset with fewer but more significant variables to ensure a good predictive model. We applied the following variable reduction techniques to reduce the dimensionality in the dataset:

- Correlation between variables
- Variation Inflation Factor analysis
- Reduction based on variance
- Feature selection from predictive model
- Transformation of Variables

1. <u>Correlation between variables</u>

The correlation between variables is an important factor in deciding the right kind of variables that should be supplied to a predictive model. A model with variables having a high correlation between them can increase the complexity of the model without improving the predictive capabilities of the algorithm. We also noticed that noise in the dataset was a major issue and hence removed variables with a high correlation of 0.7 among the variables.

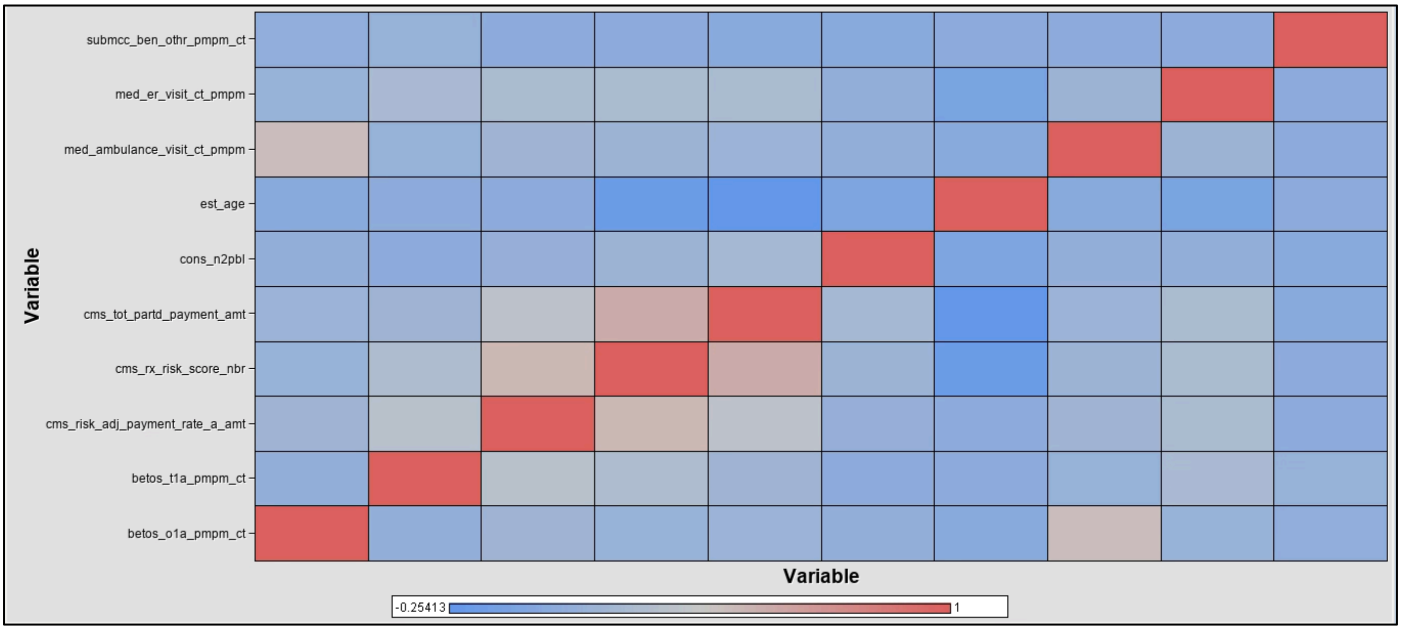

Figure 5

There were also certain indicator variables which had redundancy. For example, the data in the Berenson-Eggers Type of Service (BETOS) codes had a binary flag and a numeric value for the number of times a member submitted a claim for a BETOS code. Since there was no need to have both the binary indicator and the numeric value, certain indicators were removed from the dataset because of redundancy in information.

2. Variation Inflation Factor Analysis

Features that have a high multicollinearity between variables can significantly impact the results in predictive models. The variation inflation factor (VIF) performs a regression-based technique to eliminate variables that have a high VIF factor. Variables that had a VIF > 10 were removed from the model and not used as predictor variables. This technique helped reduce close to 35 variables.

3. Reduction Based on Variance

Categorical variables with too many levels or high variance were removed from the model. For example, variables like the state and county names were not included in the model since they would have a negative impact on the predictive capabilities of the model. Similarly, certain categorical variables like indicator variables that had 90% in a class were removed since they did not show any variance which could be captured and trained for the machine learning algorithm.

4. Feature selection from predictive model

Reducing the number of features was still our most important process before performing any modelling to predict the transportation issues in members. With an imbalanced dataset, plugging all the features into the model would result in a garbage model. In the world of machine learning there is a famous saying, “garbage in, garbage out” or the GIGO. This basically indicates the importance of feature selection. We used various tree-based machine learning algorithms and ensemble models to narrow does the feature list based. Gradient boosting and random forest algorithms were used to select the top features based on variable importance

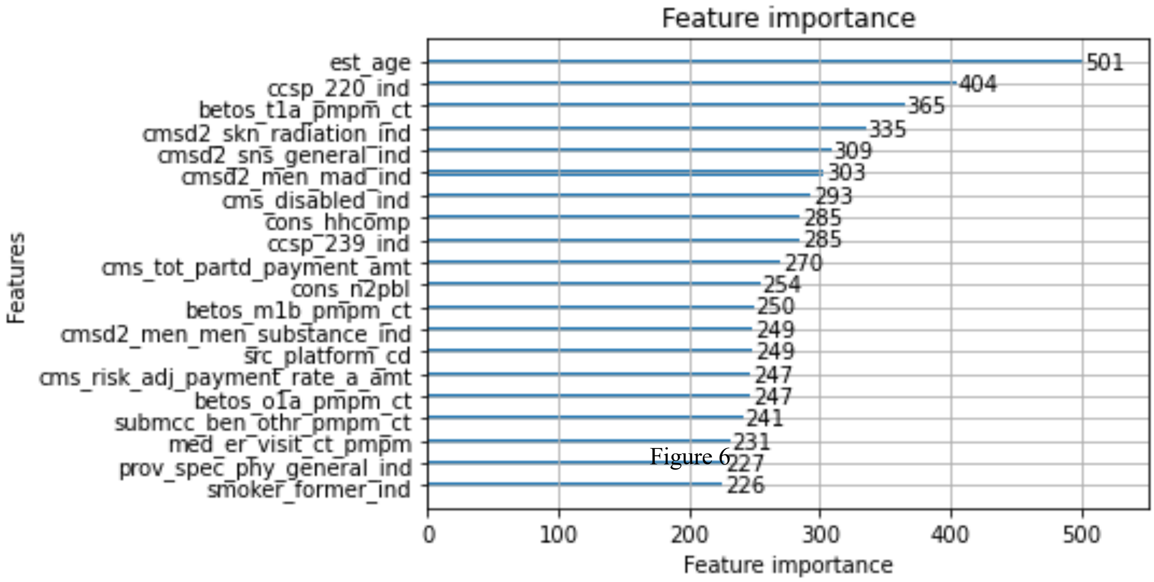

Figure 6

5. Transformation of Variables

Certain variables needed special transformation techniques to balance the skewness and kurtosis exhibited in the data. Different kinds of transformation techniques like log transformation, power series transformation and square root transformation techniques were used to model the data. Below is an example of one such variable where transformation was used. The age of the members in the data can be seen in the image on the left and the transformed variable is on the right.

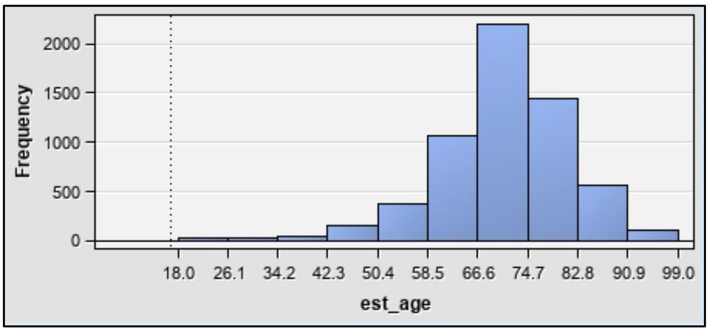

Figure 7

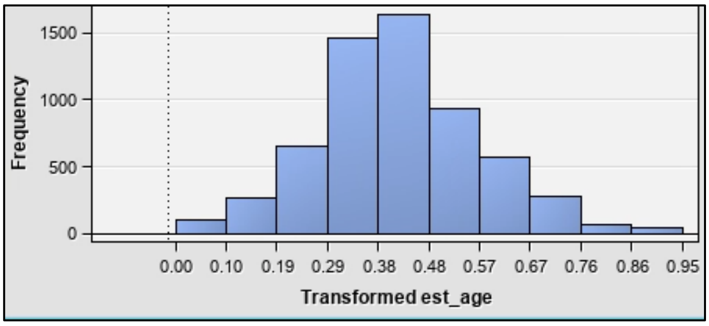

Figure 8

Predictive Analysis

The main objective for this project was to create a predictive model that can use the features in the provided dataset to identify members that have a transportation issues and cannot get access to medical care or any other life dependent services. Since the transportation issues was the dependent variables and classified as a binary indicator (1= transportation issue & 0 = no transportation issue), the problem was a perfect classification problem and needed a robust classification algorithm to handle the various characteristics in the issue. One of the main issues with the dataset was the class imbalance in the dependent variable where only 14% of the dataset was classified as a transportation issue. This was a classic case of causing a machine learning model to overfit if no prior data sampling techniques were applied during modelling. We looked at this class imbalance issues and decided to use a SMOTE based approach where the minority class (transportation issue =1) would be up sampled to reduce the imbalance in training data.

Variable Name	Description
est_age	Member age calculated using est_bday, relative to score/index date
ccsp_220_ind	Binary indicator for CCS code - Intrauterine hypoxia and birth asphyxia
betos_t1a_pmpm_ct	Per Member Per Month Count of Logical Claims for each of the BETOS codes
cmsd2_skn_radiation_ind	Binary indicator for each of the CMS Level 2 diagnosis categories present in the reference table
cmsd2_sns_general_ind	Binary indicator for each of the CMS Level 2 diagnosis categories present in the reference table
cmsd2_men_mad_ind	Binary indicator for each of the CMS Level 2 diagnosis categories present in the reference table
cms_disabled_ind	Disability Indicator
cons_hhcomp	KBM-Category-Household Composition
ccsp_239_ind	Binary indicator for CCS code - Superficial injury, contusion
cms_tot_partd_payment_amt	Total PartD Payment Amount
cons_n2pbl	KBM-Census % Black
betos_m1b_pmpm_ct	Per Member Per Month Count of Logical Claims for each of the BETOS codes
cmsd2_men_men_substance_ind	Binary indicator for each of the CMS Level 2 diagnosis categories present in the reference table
src_platform_cd	Standard Humana Member Identifier - unique when paired with src_mbr_id

cms_risk_adj_payment_rate_a_amt	CMS Risk Adjustment Payment Rate A
betos_o1a_pmpm_ct	Per Member Per Month Count of Logical Claims for each of the BETOS codes
submcc_ben_othr_pmpm_ct	Per Member Per Month Count of logical claims for each of the MCC categories present in the reference table
med_er_visit_ct_pmpm	Per Member Per Month Visits for non-BH related claims, broken out by utilization category
prov_spec_phy_general_ind	Binary indicator for a select group of categories using std_hipaa_prov_spec_cd
smoker_former_ind	Former Smoker based on the presence of smoking indication from membership or medical claims data
transportation_issues	1 = transportation challenge, 0 = no transportation challenge

Table 4: Description of important variable

Various machine learning algorithms like Light Gradient Boosting, Random Forest, XGBoost, Neural networks and SVM were used to model the data. Random Search techniques were used to tune the hyperparameters of these models and a 5-fold cross validation was performed to reduce the model to overfit on the training dataset. The data was also split into a 70-30 training validation split to understand the performance of the model on unseen data. The models developed had an accuracy of 80% but the main selection criterion for the models were based on the area under the receiver operating characteristic curve (ROC-AUC). Below is a graph highlighting the ROC index for various model on the training and validation data.

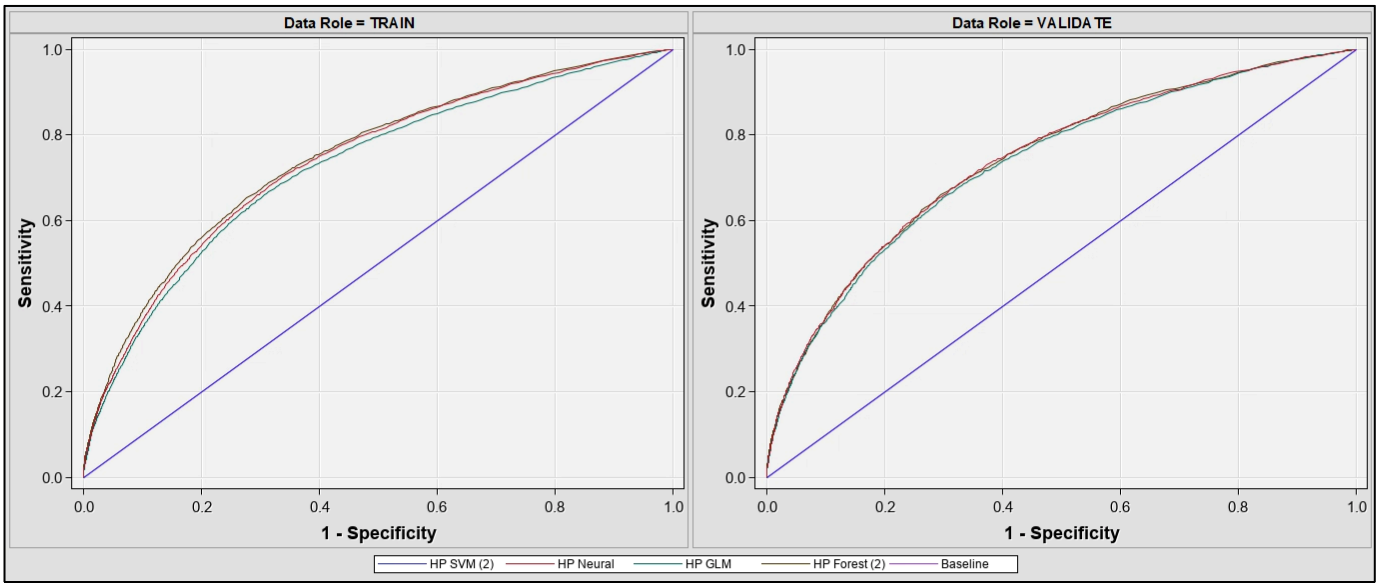

Figure 9

Results

Based on the various predictive models implemented, the Light GBM model was selected as the best model based on the AUC statistic. The model had a ROC-AUC of 0.74 on the validation data which was the best among all the models

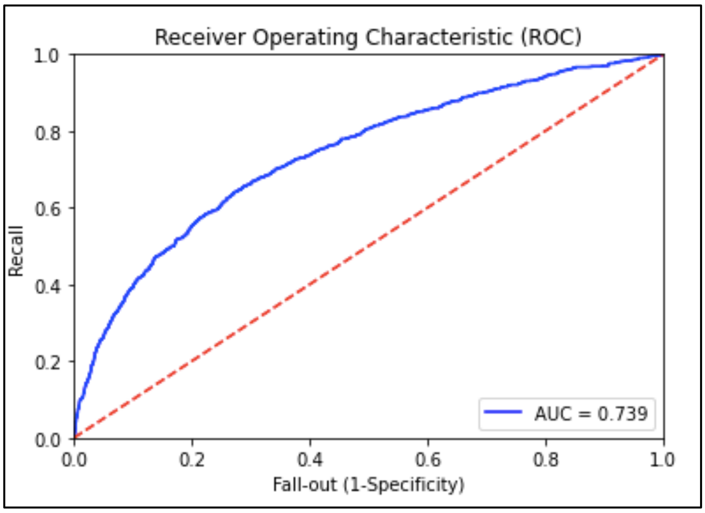

Figure 10

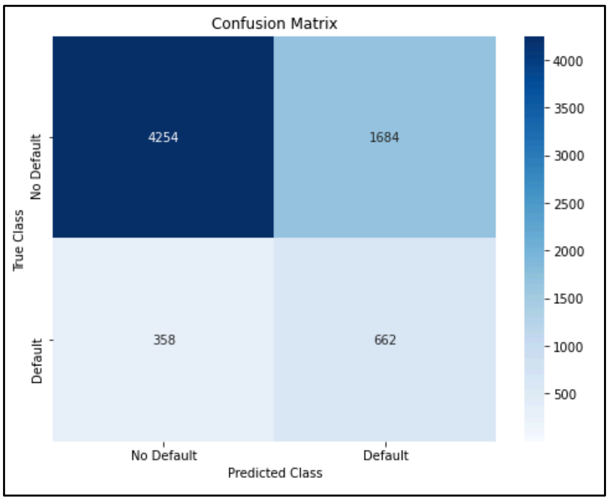

Figure 11

From the above confusion matrix, we can see that the model was able to classify 662 cases as true positives and 4254 cases as true negatives. But there were certain false positive and false negatives which could not be predicted correctly, and this had an impact on the results of the model.

Recommendations

Based on our model, we were able to create 3 patient segments. The 3 segments are equally distributed and are similar within each segment and are different from each other.

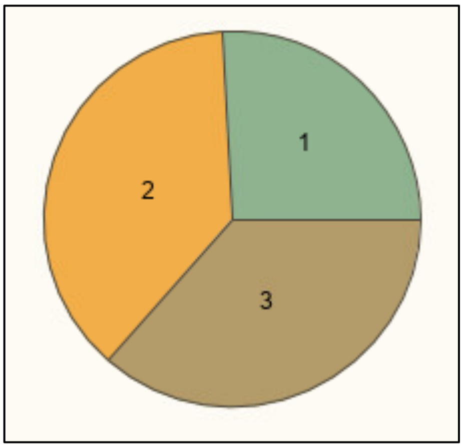

Figure 12

This customer segments are based on the following variables (in the order mentioned):

1. NP/PA assistance required (Y/N)
2. Smoking indicator (Former and current)
3. Sex
4. Language Spoken
5. Low income indicator
6. Disability indicator
7. Household composition

The 3 segments can be characterized as follows based on the severity of the transportation issue:

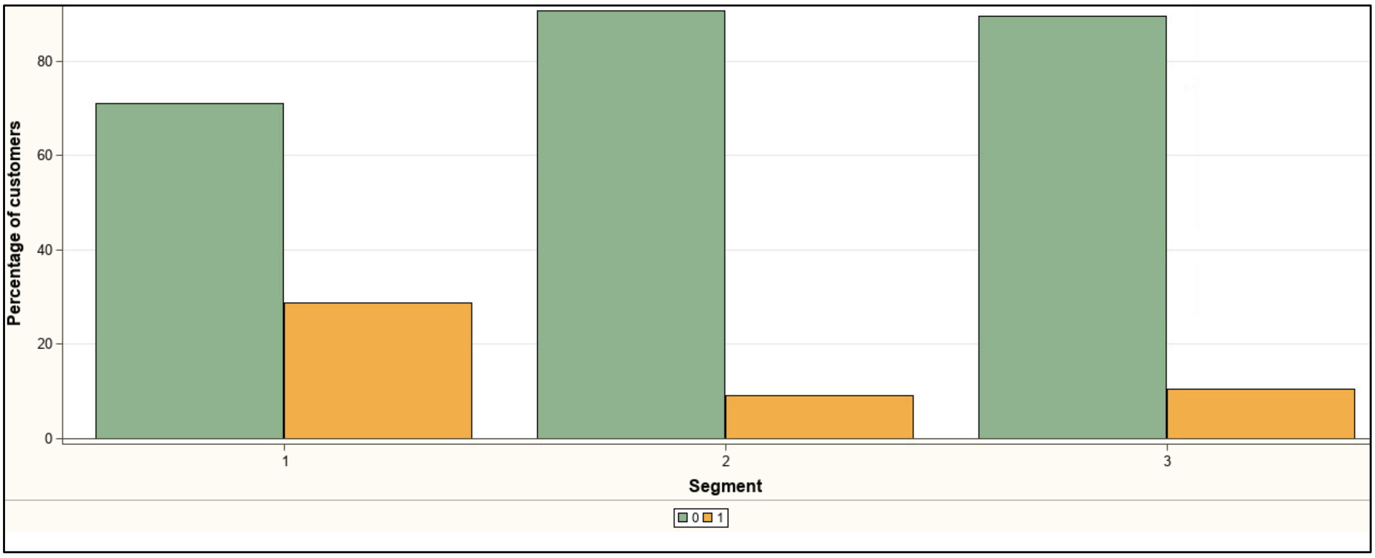

Figure 13

Segment 1: Classified as 'High Severity' patients

Segment 3: Classified as 'Medium Severity' patients

Segment 2: Classified as 'Low Severity' patients

Variables which differentiate each segment (in order of importance):

High Severity Patients	Medium Severity Patients	Low Severity Patients
Sex	NP/PA Assistance Indicator	Language Spoken
Disability Indicator	Sex	
Household composition	Disability Indicator	
Intrauterine hypoxia and birth asphyxia Indicator	Intrauterine hypoxia and birth asphyxia Indicator	
Low income indicator		

Table 5: Important variables for customer segmentation

The 3 categories can be profiled using the following variables:

Characteristics	High Severity	Medium Severity	Low Severity
Severity of transportation Issue	28%	10%	9%
Sex (Female)	64%	32%	82%
Language Spoken (Spanish)	2%	7%	5%
Disability Indicator	88%	12%	9%
Low Income Indicator	65%	8%	12%
NP/PA Assistance Indicator	60%	14%	65%
Current smokers	28%	10%	9%

Table 6: Profiles of each clusters created

Based on the above analysis of customer segment groups and research conducted to understand the impact of transportation issues as a social determinant of health, we have identified few measures that can be taken up by Humana Inc. to reduce this impact:

⇒ Research has shown that transportation barriers is one of the most significant social determinants for health and every year around 3.6 million people in the U.S do not obtain medical care due to these barriers. Therefore, it is very important to identify members who are more prone to facing such transportation challenges. From the results of our clustering, we can see that members in the "High severity" segment typically is female, have some form of disability, have a low income and need advanced healthcare professionals. We recommend conducting regular surveys and information gathering from members on a regular basis to understand these socio-economic variables that can help categorize policy members. Surveys similar to the CDC Transportation Health Impact Assessment Toolkit is a suitable starting point in helping build such surveys. The data from these surveys need to be accurate and periodic to obtain actionable insights.

⇒ Clusters based on zip codes need to be identified to highlight areas where there is a lack of public transport and affordable transportations systems. For example, members living in rural parts of the town with low incomes would not be able to travel long distances to get suitable medical care and might miss periodic medical appointments and prescription fill ups which can aggregate a disease and eventually increase the medical costs.

⇒ Transportation barriers impact not just patients but hospital and health care systems. Patients missing doctors' appointments and failing to refill on prescription drugs could eventually reduce the business and also increase the medical costs for patients. Therefore, we recommend Humana Inc. can tie up with local hospital and medical centers to create a network to offer a more flexible policy for members to schedule appointments and waive off certain fees associated with cancelling or modifying appointments.

⇒ Data is going to play a crucial role in trying to solve transportation issues for members. Humana Inc. needs to invest heavily in recording, storing and analyzing information from various members to be able to make data driven decisions. Our profiling statistics has shown patients with a low income to be at a higher risk of facing transportation barriers and the Bureau for Labor Statistics also shows that people earning between $5,000 and $30,000 per year spend 24% of their income on transportation. Hence, it is critical that Humana Inc. leverages the data that is collected from individuals to categorize them into the right categories to prevent an unwanted increase in medical costs due to transportation issues.

⇒ Partnerships with the government, local bodies and non-profit organizations is going to be a crucial factor in helping reduce the burden of transportation issues as a social determinant to health. Humana can partner with various non-profit organizations and ride share companies to provide more affordable transportation options for individuals to access better medical care facilities. For example, allowing members to use Uber to visit hospital within the Humana network can be a win-win for both Humana, Uber and the hospitals within the Humana network.

Conclusion

We hope with the efforts that Humana Inc. is currently putting into solving transportation issues as social determinants to health will have a huge impact on the way citizens live and manage their health. With the advancements in science and technology, problems like this will soon find a solution and the society will greatly benefit from these changes.

Appendix

List of tables:

List of figures: