Humana-Mays Healthcare Analytics 2020 Case Competition

Identifying Medicare Members with Transportation Challenges

Social determinants of health (SDoH) are becoming more and more important in maintaining the overall health of human-beings. Transportation accessibility, which is fundamental for individuals' need to engage with their community, is one of them. As such, accurately predicting whether or not someone will be facing transportation challenges opens many opportunities for healthcare companies to better know about its members, not only helping those in need but also helping them in advance. Using one-year longitudinal data with 800+ features, we built a deep neural network model after applying natural language processing, which could predict whether a Medicare member will encounter a transportation issue with an AUC of .7406. In addition to prediction, our model offers interpretable insights that allow us to formulate recommendations related to sharing actionable future steps with Humana.

Contents

1. Introduction

Social determinants of health (SDoH) are the conditions in which people are born, grow, live, work and age that shape health[1]. To most of us, it seems like medical care or health care plays a crucial role in maintaining the health to human-beings, whereas it is not the case. Medical care only produces 10 percent of what creates health, with 60 percent of those relating to the interplay between our socio-economic and community environments and lifestyle behaviors, aka SDoH[2]. Therefore, addressing social determinants of health is important for improving health and reducing health disparities[3].

Humana Inc., a leading healthcare company is now seeking to better understand the 'total health' of its members by exploring factors in SDoH thus creating improved access to health resources and service, such as healthy food and beverage, safe home and community environments, economic opportunity, education and retraining, etc.

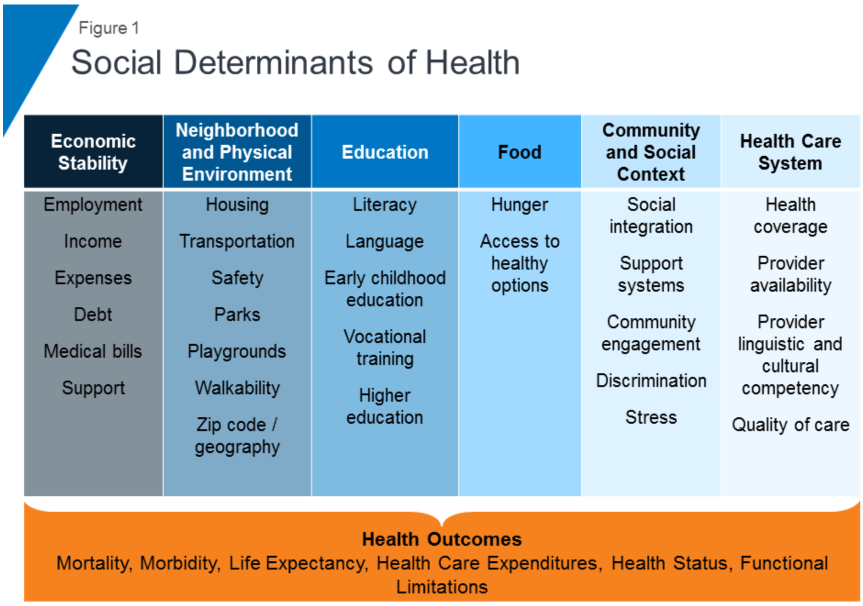


Figure 1

Social Determinants of Health

Economic Stability	Neighborhood and Physical Environment	Education	Food	Community and Social Context	Health Care System
Employment Income Expenses Debt Medical bills Support	Housing Transportation Safety Parks Playgrounds Walkability Zip code / geography	Literacy Language Early childhood education Vocational training Higher education	Hunger Access to healthy options	Social integration Support systems Community engagement Discrimination Stress	Health coverage Provider availability Provider linguistic and cultural competency Quality of care

Health Outcomes

Mortality, Morbidity, Life Expectancy, Health Care Expenditures, Health Status, Functional Limitations

Figure 1: Social Determinants of Health

As one of the important SDoH (as shown in figure 1), transportation accessibility is fundamental for individuals' need to engage with their community, for obtaining employment, goods and services, health, and education, and for socializing. Thus, identifying potential transportation challenge and offering timely and useful solutions contribute to the fulfillment of 'total health'.

The aim of our analysis is to dig deep into the transportation area, exploring related factors

[1] "About Social Determinants of Health," World Health Organization, accessed April 25, 2018, http://www.who.int/social_determinants/sdh_definition/en/.

[2] "Making a down payment on health by creating social," Kaiser Permanente, accessed November 4, 2015, https://about.kaiserpermanente.org/community-health/news/making-a-down-payment-on-health-kaiser-permanente-invests-in-cre

[3] "Healthy People 2020: Social Determinants of Health," Office of Disease Prevention and Health Promotion, accessed April 25, 2018, https://www.healthypeople.gov/2020/topics-objectives/topic/social-determinants-of-health.

which could lead to transportation challenges, and develop a classification model to identify and predict Medicare members who are most likely struggling with transportation issues. Most importantly, we want to provide feasible recommendations and generate actionable insights according to different segments of members to inform Humana's business decisions.

2. Data Preparation

2.1 General Data Cleaning

2.1.1 Missing Value Processing

131 columns (features) in the training dataset have missing values. After initial checking, we dealt with them in three ways respectively.

- Drop the feature if most of the values are missing.
- Replace the missing value with the mode if it's a discrete feature.
- Replace the missing value with the mean if it's a continues feature.

There are other columns whose values have special symbols (e.g. column 'cms_ra_factor_type_cd' has '*') or repeated values in different forms (e.g. column 'lang_spoken_cd' has 'E' which means 'ENG'). We treated them respectively as well.

- Replace the special symbol with the mode if the feature DOES NOT have an unknown category before.
- Replace the special symbol with 'U' or similar representations if the feature HAS an unknown category before.
- Replace the repeated value with the one has the same meaning but different form.

After all the missing value processing work, three columns have been dropped (hedis_ami, hedis_cmc_ldc_c_control, hedis_cmc_ldc_c_screen) and others have been replaced properly.

2.1.2 Single-Valued Features Dropping

Since it is unnecessary to include features which only have one unique value, we decided to drop all of these 95 columns below. Most of them are from MCC category.

'ccsp_034_ind'	'med_ip_maternity_admit_days_pmpm'
'ccsp_120_ind'	'med_ip_mhsa_admit_ct_pmpm'
'hlth_pgm_slvrsnkr_refer_status'	'med_ip_mhsa_admit_days_pmpm'
'lab_hba1_c_abn_result_ind'	'rev_cms_nicu_ind'
'med_ip_ltach_admit_ct_pmpm'	'rx_gpi2_08_pmpm_ct'
'med_ip_ltach_admit_days_pmpm'	'rx_gpi2_14_pmpm_ct'
'med_ip_maternity_admit_ct_pmpm'	'rx_gpi2_19_pmpm_ct'
'rx_gpi2_20_pmpm_ct'	'rx_gpi2_45_ind'
'rx_gpi2_29_pmpm_ct'	'rx_gpi2_84_ind'
'rx_gpi2_45_pmpm_ct'	'submcc_pre_del_ind'
'rx_gpi2_69_pmpm_ct'	'submcc_rar_drm_ind'
'rx_gpi2_70_pmpm_ct'	'rx_gpi2_81_ind'
'rx_gpi2_76_pmpm_ct'	'rx_gpi2_29_ind'

'rx_gpi2_80_pmpm_ct'	'rx_gpi2_14_ind'
'rx_gpi2_81_pmpm_ct'	'submcc_ben_lymp_ind'
'rx_gpi2_84_pmpm_ct'	'rx_gpi2_19_ind'
'rx_gpi2_92_pmpm_ct'	'submcc_pre_othr_ind'
'rx_gpi2_95_pmpm_ct'	'submcc_rar_othr_ind'
'rx_gpi2_96_pmpm_ct'	'rx_gpi2_92_ind'
'rx_gpi2_98_pmpm_ct'	'rx_gpi2_95_ind'
'submcc_ben_lymp_pmpm_ct'	'submcc_hdz_surg_ind'
'submcc_brn_acc_pmpm_ct'	'submcc_cad_fh/ho_ind'
'submcc_cad_fh/ho_pmpm_ct'	'submcc_neo_fh/ho_ind'
'submcc_hdz_it_i_pmpm_ct'	'rx_gpi2_08_ind'
'submcc_hdz_surg_pmpm_ct'	'rx_gpi2_69_ind'
'submcc_hiv_kapo_pmpm_ct'	'submcc_rar_als_ind'
'submcc_hiv_pcp_pmpm_ct'	'submcc_rar_cf_ind'
'submcc_inf_men_pmpm_ct'	'submcc_rar_sca_ind'
'submcc_neo_fh/ho_pmpm_ct'	'submcc_inf_men_ind'
'submcc_pre_del_pmpm_ct'	'submcc_hdz_it_i_ind'
'submcc_pre_ect_pmpm_ct'	'submcc_trm_fxu_ind'
'submcc_pre_mul_pmpm_ct'	'submcc_hiv_pcp_ind'
'submcc_pre_othr_pmpm_ct'	'submcc_rar_pol_ind'
'submcc_rar_als_pmpm_ct'	'submcc_rsk_fh/h_ind'
'submcc_rar_cf_pmpm_ct'	'submcc_rsk_pcos_ind'
'submcc_rar_drm_pmpm_ct'	'submcc_pre_ect_ind'
'submcc_rar_othr_pmpm_ct'	'rx_gpi2_76_ind'
'submcc_rar_pol_pmpm_ct'	'rx_gpi2_80_ind'
'submcc_rar_sca_pmpm_ct'	'submcc_pre_mul_ind'
'submcc_rsk_an_pmpm_ct'	'rx_gpi2_98_ind’
'submcc_rsk_fh/h_pmpm_ct'	'rx_gpi2_20_ind'
'submcc_rsk_othr_pmpm_ct'	'submcc_brn_acc_ind'
'submcc_rsk_pcos_pmpm_ct'	'submcc_rsk_an_ind'
'submcc_trm_fxu_pmpm_ct'	'submcc_hiv_kapo_ind'
'total_ip_ltach_admit_ct_pmpm'	'rx_gpi2_70_ind'
'total_ip_ltach_admit_days_pmpm'	'rx_gpi2_96_ind'
'total_ip_maternity_admit_ct_pmpm'	'submcc_rsk_othr_ind'
'total_ip_maternity_admit_days_pmpm	

2.2 Data Manipulation For Different Models

After initial data cleaning, we decided to further manipulate data according to different model requirements (We will further demonstrate the reason for using the following two models in chapter 3).

For DeepFM model, we need to first reverse one-hot features, then divided continuous features into categories, and lastly put all the processed features into fields.

For Deep Neural Network model, we finally decided to use Natural Language Processing technique to manage data after several other attempts.

2.2.1 DeepFM

First, we wanted to reduce the dimension of the dataset by reversing one-hot features to build categorical features. Unfortunately, only one set of two features is one-hot encoded, "cons_hcaccprf_c" and "cons_hcaccprf_p". According to the data documentation, "cons_hcaccprf" represents a member's healthcare treatment preference modeled by KBM, and one type "O – Other or None Above" is included. So, we built a new feature called "cons_hcaccprf" and made the value to be "c" or "p" if "cons_hcaccprf_c" or "cons_hcaccprf_p" equals one, then we filled the missing values with "o". There are still a lot of binary features in the dataset but they all represent different fields of information, so our dimension reduction work had to stop here.

Second, to deal with unbalanced dense features and minimize the influence of outliers, we divided continuous features into categories. Considering that different features had different meanings, our method of binning varied. Most of the continuous features belonged to Medical Claims Features, Credit data, and Pharmacy Claims Features. Demographics/Consumer data also included some basic continuous features like "est_age".

To be more specific, when we dealt with features like "est_age" that had a limited range and nearly symmetrical distribution, we used quartiles to group the values. Take "est_age" for instance, the middle 50% of members were between 66 and 77 years old. We used the quartile 1 (66) and 3 (77) to calculate the IQR and then we found out that members who were below 50 years old and above 92 years old were outliers, so we divided these two types of observations into unique groups. Then we set three lines using quartile 1 (66), median (71), and quartile 3 (77) to divide the rest values into four groups (left inclusive): 50-66, 66-71, 71-77, 77-92. After binned, "est_age" had 6 groups. There were more than 20% of observations in each main group and the proportion of outliers was below 5%.

We also used this kind of binning method to deal with some Credit data. Features containing auto loan information mainly had similar distributions with what we discussed above, so we treated them using the same way as above.

Other continuous features (mainly in Medical Claims data and Pharmacy Claims data) were seriously right skewed and a significant number of observations were near 0. In this condition, the 3 quartiles were very close so we did not use quartiles to group data. Instead, we used equal width binning, dividing data into 5 equal width groups.

Besides, we used a special method to deal with "betos_xxx_pmpm_ct" series features. these features included information about per member per month count of logical claims for each of the BETOS codes. Because these values were calculated using total counts divided by the number of months that the member stayed with Humana during the past 12 months before the survey date, we multiplied the values with 12 first. Then the values could be treated like the reciprocal of the frequency. We grouped the values to categories like "once per year", "twice per year", "once per quarter", "once per month", "twice per month" and "more than twice per month". In this way, we also transformed the rightly skewed continuous features to balanced categorical features.

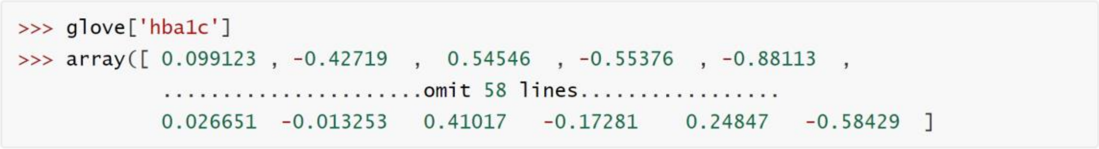
Third, we divided these features into fields according to the information they contain.

In general, we created 8 fields based on the information in the kickoff presentation, Medical Claims Features, Pharmacy Claims Features, Lab Claims Features, Demographics, Credit data, Condition Related Features, CMS Features, and Other Features.

2.2.2 Natural Language Processing

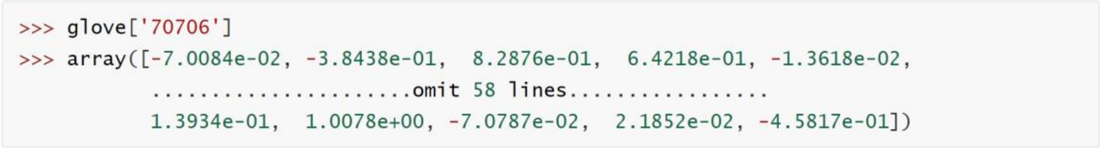
We considered that there might be some medical information behind value names which may not be represented in the dataset. So, we tried to add literal meaning to the dataset.

We used a pre-trained set of Global Vectors for Word Representation (GloVe) to convert each medical word to a 300 dimensions vector. For example:



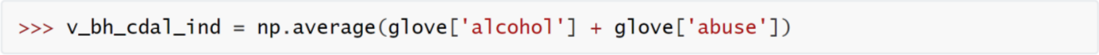
```
>>> glove['hba1c']
>>> array([ 0.099123 , -0.42719  ,  0.54546  , -0.55376  , -0.88113  ,
          ......................omit 58 lines................
          0.026651  -0.013253   0.41017   -0.17281    0.24847   -0.58429   ]
```

What needs to be mentioned is that GloVe also contains geographic information such as zip code:



```
>>> glove['70706']
>>> array([-7.0084e-02, -3.8438e-01,  8.2876e-01,  6.4218e-01, -1.3618e-02,
         ......................omit 58 lines................
         1.3934e-01,  1.0078e+00, -7.0787e-02,  2.1852e-02, -4.5817e-01])
```

If a variable name contains more than one word, the vector should be the average of the vectors of its words. For example,



```
>>> v_bh_cdal_ind = np.average(glove['alcohol'] + glove['abuse'])
```

- For indicator variables, its final value is its word-vector multiply its original value (either 0 or 1). That is to say, we either keep its word-vector or set it all zero.
- For indicator variables with per member per month count, in addition, we multiplied its paired count value.
- For other spares features, we directly replaced the original value with its word-vector.
- For all dense features, we simply replaced the original value with the product of itself and its word-vector.

The we used a deep neural network to fit the converted dataset.

3. Predictive Modeling Approach

3.1 Model Requirements

After cleaning the dataset and building features, we got a wide dataset containing 722 features. To begin our work, we split our dataset into two parts, a training dataset, and a test dataset. The training dataset contained 80% of the observations and the test dataset contained the remaining 20%. We used the training dataset to train models, then we evaluated and compared different models using the test dataset.

We began by evaluating the training data to see what kind of model was optimal to do binary classification on the data. We found that the data had several important characteristics:

First, the data is very wide. The original data had 826 columns. After cleaning the data, we kept 722 features and 1 target column. High dimensionality influences the model training process somehow because there is a high probability that each row has some outliers in some dimensions.

Second, the data lacks a direct relationship. All the 722 features have very low correlations with the target column. A correlation of 0.15 is relatively high in the dataset. So, the ability to find hidden information is important. The model needs to have a strong ability in building high-level features.

Third, many of our features are highly correlated. In general, features that belong to the same field are more likely to have the problem. In the Credit data field and Pharmacy Claims Features field, the collinearity problem is most severe. Credit data includes features that represent different parts of the same problem. For example, as to a member's balance of all credit type, the dataset provides 4 different features according to different days past due. In conclusion, the dataset includes a great number of features that are naturally correlated.

Fourth, the value of the target column is highly unbalanced. Both in the training dataset and test dataset, the proportion of observations who reported having "transportation issue" to those who reported not is around 1:6. This is a serious problem, which means the model needs to learn deep inside the dataset to become precise enough. Besides, simple tree models for classification are likely to be overfitted facing this problem.

Fifth, continuous features have different ranges and scales. Because the dataset contains features of many aspects of observations, including their financial condition and the dose of drugs they use, the diversity of ranges in data is natural. For example, "rx_gpi2_xx_pmpm_ct" is a series of continuous features that have a very limited range. Most features have a max value of 2 and some features even range between 0 and 0.2. But in credit data, observations' loan balances can reach 100000.

3.2 Final Model (Deep Neural Network after Natural Language Processing)

In this section, we first discussed how we fit DNN model with the original one-hot featured training data. Then we adopted NLP to process the data and fit DNN model to it again. We compared the AUC for these two methods and found that the performance of the latter one is much better.

3.2.1 DNN with One-hot Featured Dataset

A deep neural network is a neural network with more than two layers. The input of each layer depends on the output of its previous layer. Theoretically, a deep neural network is able to fit any nonlinear function. The selling point of a deep neural network is that it can automatically compute high-order cross-products, liberate human labor from doing future engineering manually. If fine-tuned properly, a deep neural network can produce a relatively good performance.

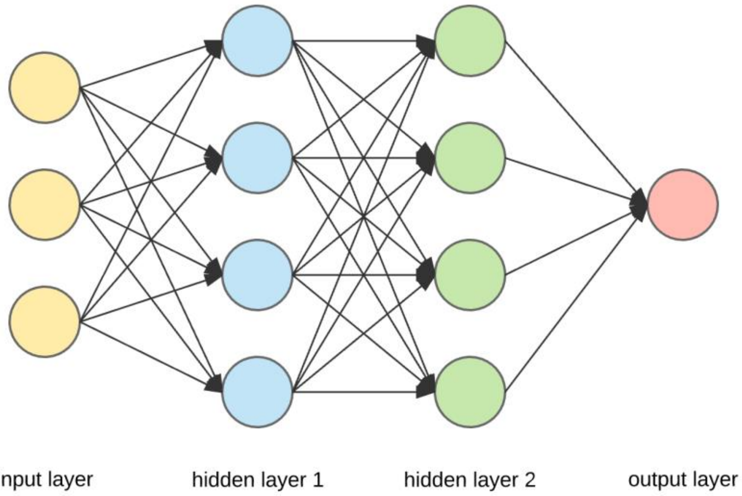
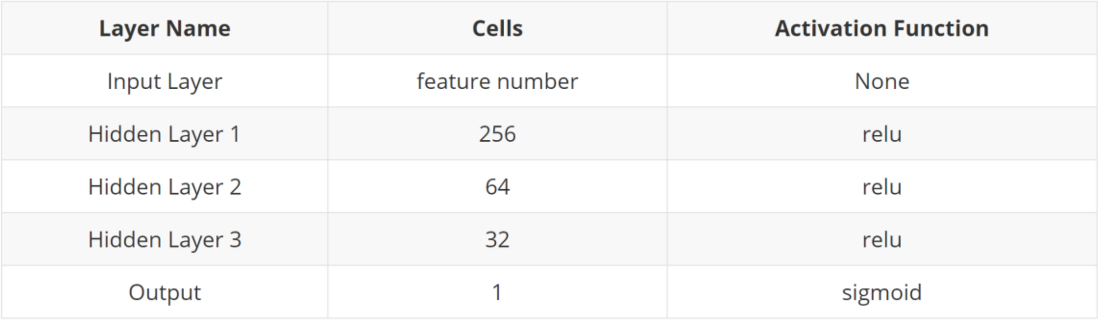


Figure 2: A Deep Neural Network Example

During the fine-tuning process, we need to set several hyperparameters, which are the learning rate, the number of layers, the number of cells in each layer, and the class weight due to the imbalance of samples.

Unlike tree models, a deep neural network does not give feature importance itself, so we have to derive feature importance manually by the following steps. First, we picked on variables, instead of dropping it directly, we shuffled it and put it back to the dataset. Second, we fit the new dataset with the same neural network structure. Finally, we computed the difference of AUC. The lager the difference is, the more important the variable is.

After fine-tuning, we set all learning rate as 0.01 and the network structure is shown below.

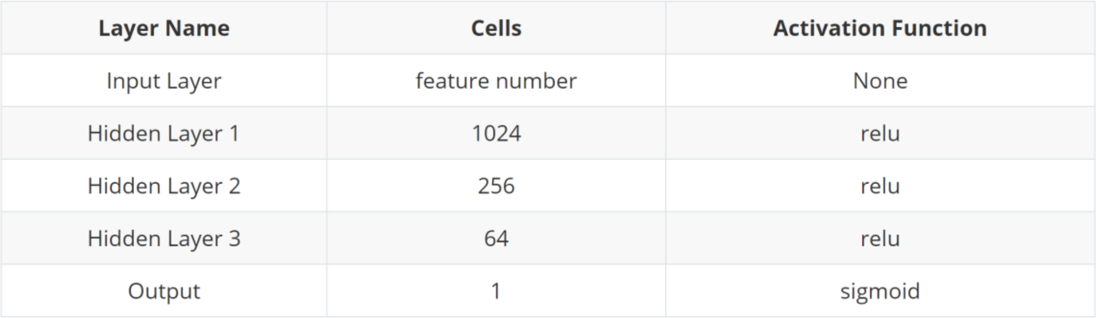


Layer Name	Cells	Activation Function
Input Layer	feature number	None
Hidden Layer 1	256	relu
Hidden Layer 2	64	relu
Hidden Layer 3	32	relu
Output	1	sigmoid

This model gets an AUC of 0.7371.

3.2.2 DNN with NLP Dataset

We managed the training data with Natural Language Processing technique (see section 2.3.2) first and then fit the DNN model to our data. After fine-tuning, we set all learning rate as 0.001 and the network structure is shown below.



Layer Name	Cells	Activation Function
Input Layer	feature number	None
Hidden Layer 1	1024	relu
Hidden Layer 2	256	relu
Hidden Layer 3	64	relu
Output	1	sigmoid

This model gets an AUC of 0.7406 which is better than the former one. Thus, we decided to adopt it at last.

3.3 Exploring Other Models

Before making our final decision that a deep neural network model using data processed by natural language processing is the most appropriate model, we investigated other classification methods that could potentially help us to interpret the results easier, or achieve higher accuracy in identifying Medical members who are struggling with transportation challenges.

3.3.1 Random Forest

Random forest is a Bagging method in integrated learning, and it's based on decision tree. Decision tree is a very simple algorithm. It is highly explanatory and conforms to human intuitive thinking. It is a supervised learning algorithm based on the if-then-else rule. Random forests are made up of many decision trees, and there is no correlation between different decision trees.

When we perform the classification task, the new input sample enters, and each decision tree in the forest is judged and classified separately. Each decision tree will get its own classification result, and which classification result of the decision tree Most, then random forest will use this result as the final result.

There are several advantages for using Random Forest:

- Random forest runtimes are quite fast, and they can deal with unbalanced and missing data.
- It can measure the importance of features.
- It is relatively simple to implement.

However, weaknesses are also obvious:

- For very large data sets, the size of the trees can take up a lot of memory.
- It can tend to overfit on datasets that are particularly noisy.

After fitting Random Forest model to our test dataset, we found that the result was not so optimistic since the AUC score turned out to be 0.63 and the recall rate for 'y=1' was below 0.2. Thus, we decided not to move on with Random Forest but trying other models.

3.3.2 Extreme Gradient Boosting

We tried Boosting models as well, which is a sequential technique that works on the principle of an ensemble. Boosting combines a set of weak learners (high bias, low variance) and delivers improved prediction accuracy. Compared to Bagging (e.g. random forest) which uses fully grown decision trees (low bias, high variance), grows trees in parallel and reduces error by reducing variance, Boosting reduces error mainly by reducing bias.

We fit XGBoost model to our training data and compared our evaluation metrics with the results of the random forest. It demonstrated that XGBoost is comparatively faster and has shown better performance, raising AUC from 0.63 to 0.67. But the recall rate for 'y=1' is still quite low, which means we need to dig deep into that segment of people further by extracting more useful information. Meanwhile, the overall AUC is not high enough, urging us to exploring other more complex models such as deep neural network models.

3.3.3 DeepFM

Since the dataset is relatively sparse, a DeepFM model seems to be applicable. The "deep" stands for deep neural network, and "FM" means factorization machine. This model emphasizes both low- and high-order feature interactions. The structure of DeepFM is shown below.

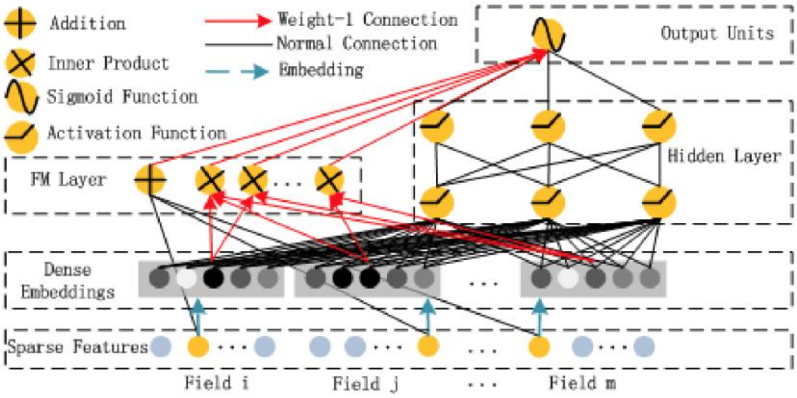
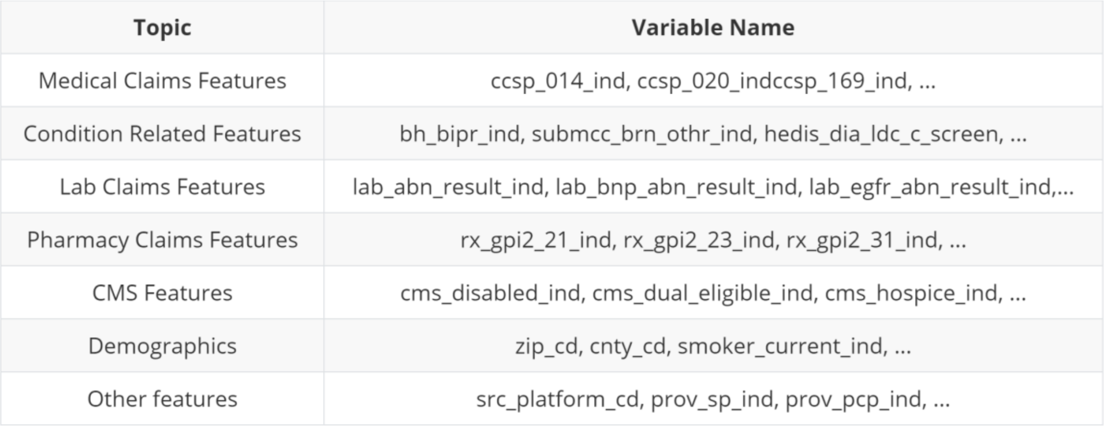


Figure 3: The Structure of DeepFM

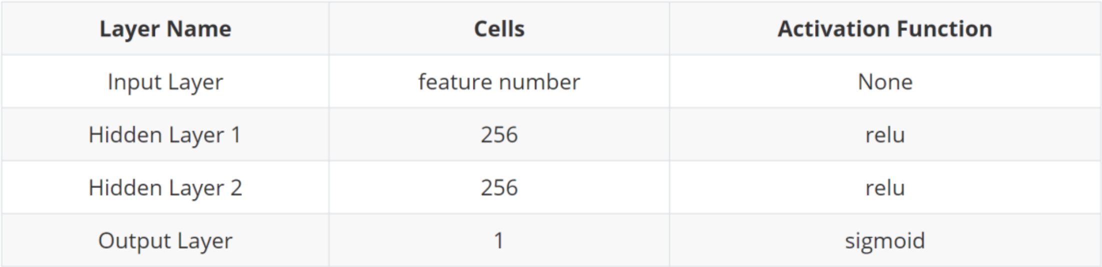
Besides all the hyperparameters mentioned in the DNN section, we also need to assign an embedding dim for all the fields. With an embedding layer, we are able to convert spares features to dense ones.

We classify all spares features into 7 topics.

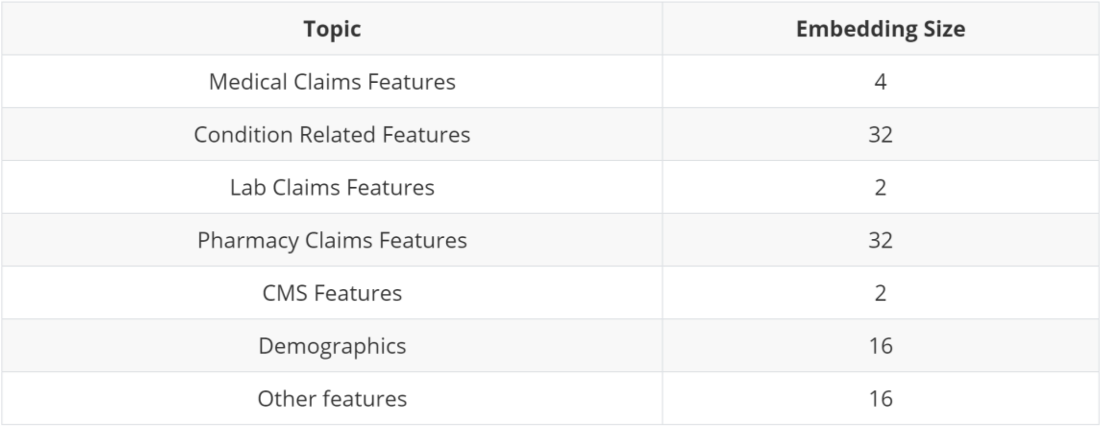


Topic	Variable Name
Medical Claims Features	ccsp_014_ind, ccsp_020_indccsp_169_ind, ...
Condition Related Features	bh_bipr_ind, submcc_brn_othr_ind, hedis_dia_ldc_c_screen, ...
Lab Claims Features	lab_abn_result_ind, lab_bnp_abn_result_ind, lab_egfr_abn_result_ind,...
Pharmacy Claims Features	rx_gpi2_21_ind, rx_gpi2_23_ind, rx_gpi2_31_ind, ...
CMS Features	cms_disabled_ind, cms_dual_eligible_ind, cms_hospice_ind, ...
Demographics	zip_cd, cnty_cd, smoker_current_ind, ...
Other features	src_platform_cd, prov_sp_ind, prov_pcp_ind, ...

After fine-tuning, we set all learning rate as 0.01 and the network structure is shown below. The embedding size of each topic are shown below.



Layer Name	Cells	Activation Function
Input Layer	feature number	None
Hidden Layer 1	256	relu
Hidden Layer 2	256	relu
Output Layer	1	sigmoid



Topic	Embedding Size
Medical Claims Features	4
Condition Related Features	32
Lab Claims Features	2
Pharmacy Claims Features	32
CMS Features	2
Demographics	16
Other features	16

This model gets an AUC of 0.717.

4. Model Evaluation

4.1 Model Metrics

The main metric of interest for this case competition — as defined by Humana — is the area under the curve (AUC). In this section, we will discuss our model's performance on our test set in terms of AUC as well as other evaluation metrics that we felt were important to assessing the quality of our model.

4.1.1 *Establishing Baseline Model Performance*

Before evaluating the performance of any modeling approach, it is critical to establish a baseline model for comparison. Baseline models are often naive modeling approaches that establish a lower bound for model performance. In this prediction problem, a baseline model could be to simply predict the outcome that is most prevalent in the training data.

However, baseline model may not suit for our current case. Because our training data is so imbalanced with a ratio of about 1 over 6 for the amount of transportation_issues = 0 compared to transportation_issues = 1. Thus, we decided not to use accuracy as a valid model metric for model evaluation.

4.1.2 *AUC*

The area under a Receiver Operating Characteristic (ROC) curve, abbreviated as AUC, is a single scalar value that measures the overall performance of a binary classifier. The AUC value is within the range [0.5-1.0], where the minimum value represents the performance of a random classifier and the maximum value would correspond to a perfect classifier (e.g., with a classification error rate equivalent to zero).

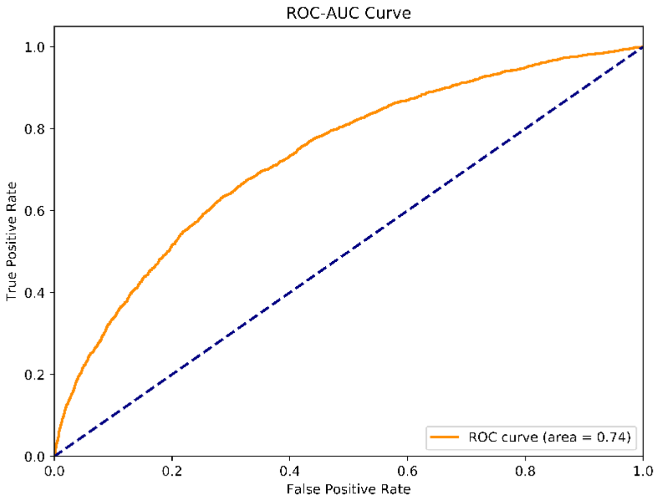


Figure 4: ROC-AUC Curve

The curve above represents the ROC of our model on the split test data. As we can see, the curve is close to the upper left corner, which indicates a high predictive power. It is significant to visualize the ROC for understanding the trade-off between the true positive rate and false positive rate.

Our final model had a .7406 AUC on our split test data. This strong AUC suggests that our model does a sound job of differentiating transportation_issues = 0 observations from transportation_issues = 1 observations. It is important to note that AUC also has a probabilistic interpretation. An AUC of .7406 implies that given the predicted score of two Medicare members — one whose outcome is truly transportation_issues = 1, and the other whose outcome is truly transportation_issues = 0 — our model will correctly assign a higher score to the transportation_issues = 1 member 74.06% of the time. The reason why AUC does not reach much higher may be the subjectivity or other personal reasons that Medicare members responded to the given question.

4.1.3 Confusion Matrix

The table below is the confusion matrix for our test data:

	Predicted transportation_issues = 0	**Predicted transportation_issues =1**
Actual transportation_issues = 0	10737	1223
Actual transportation_issues = 1	1228	727

An interesting phenomenon in our data is the relative balance between false positives and false negatives. False positives are observations we predicted to be positive but are actually negative. False negatives are observations that we predicted to be negative but are actually positive. In our training data we had 1223 false positives and 1228 false negatives, suggesting that our model has found a decent middle ground in making tradeoffs between false positives and false negatives.

However, false negatives are likely more costly than false positives. This is because it is far worse to miss a Medicare member who will face transportation challenges — and as a result fail to deliver them resources and service — than it is to provide resources and service to a Medicare member who may not have needed them. Considering the number of false negatives are more than true positives, our model may still need further improvement.

4.1.4 Precision, Recall, and F1 Score

Precision, Recall, and F1 Scores at the aggregate and class levels are helpful tools for understanding how our model performs in more specific circumstances.

	Precision	**Recall**	**F1 Score**
transportation_issues = 0	0.89	0.90	0.90
transportation_issues = 1	0.36	0.35	0.35
Average (Total)	0.82	0.82	0.82

Precision is the ratio of system generated results that correctly predicted positive observations (True Positives) to the system's total predicted positive observations, both correct (True Positives) and incorrect (False Positives).

Recall is the ratio of system generated results that correctly predicted positive observations (True Positives) to all observations in the actual malignant class (Actual Positives).

The F1 Score is the weighted average (or harmonic mean) of Precision and Recall. Therefore, this score takes both False Positives and False Negatives into account to strike a balance between precision and Recall.

Here, our precision, recall and F1 Score are consistent with each other in each category. However, all three metrics demonstrated a better performance in class 0 than class 1. This better performance on class 0 is to be expected, since a larger proportion of our training data belonged to this class.

4.2 Variable Importance

Unlike tree models, a deep neural network does not give feature importance itself, so we have to derive feature importance manually. The way to determine the importance of a certain feature should follow these steps:

First, when we choose a feature, we cannot just drop it. Instead, we shuffle the values randomly and replace the original column in the dataset. Now we get a new dataset with the same shape as the original one.

Second, we fit the new dataset with the same neural network structure and get an AUC_new.

Finally, we compute the difference between AUC_new and AUC of the original dataset. The larger the difference is, the more important the feature is.

The figure (Figure 5) below shows the top 50 important features. MCC Diagnosis Code features make up a great proportion of the top 50 (23 out of 50). Specific GPI2 Level Prescription Utilization features and some credit features also play an important role in the dataset.

We find that “submcc_res_copd_pmpm_ct” is the most important feature in the dataset, which means chronic obstructive pulmonary disease influence the transportation issue most. Without the correct information about it, the AUC of our model will decrease by 0.242938828. But as we can see, the differences between the top 50 is very small. The importance of feature rank 50 in the list is 0.240683612, and the difference is near 0.002.

In our model, 283 features have an importance score of over 0.24, making up 39.36% of all the features we create. And the number of features whose importance is over 0.2 is exactly 300. After that, features’ importance drops significantly. The feature rank 301 only has an importance score of 0.1444, while feature rank 300 has 0.2239. Moreover, in all the 719 features, 414 features are below 0.1 and 166 features are below 0.01.

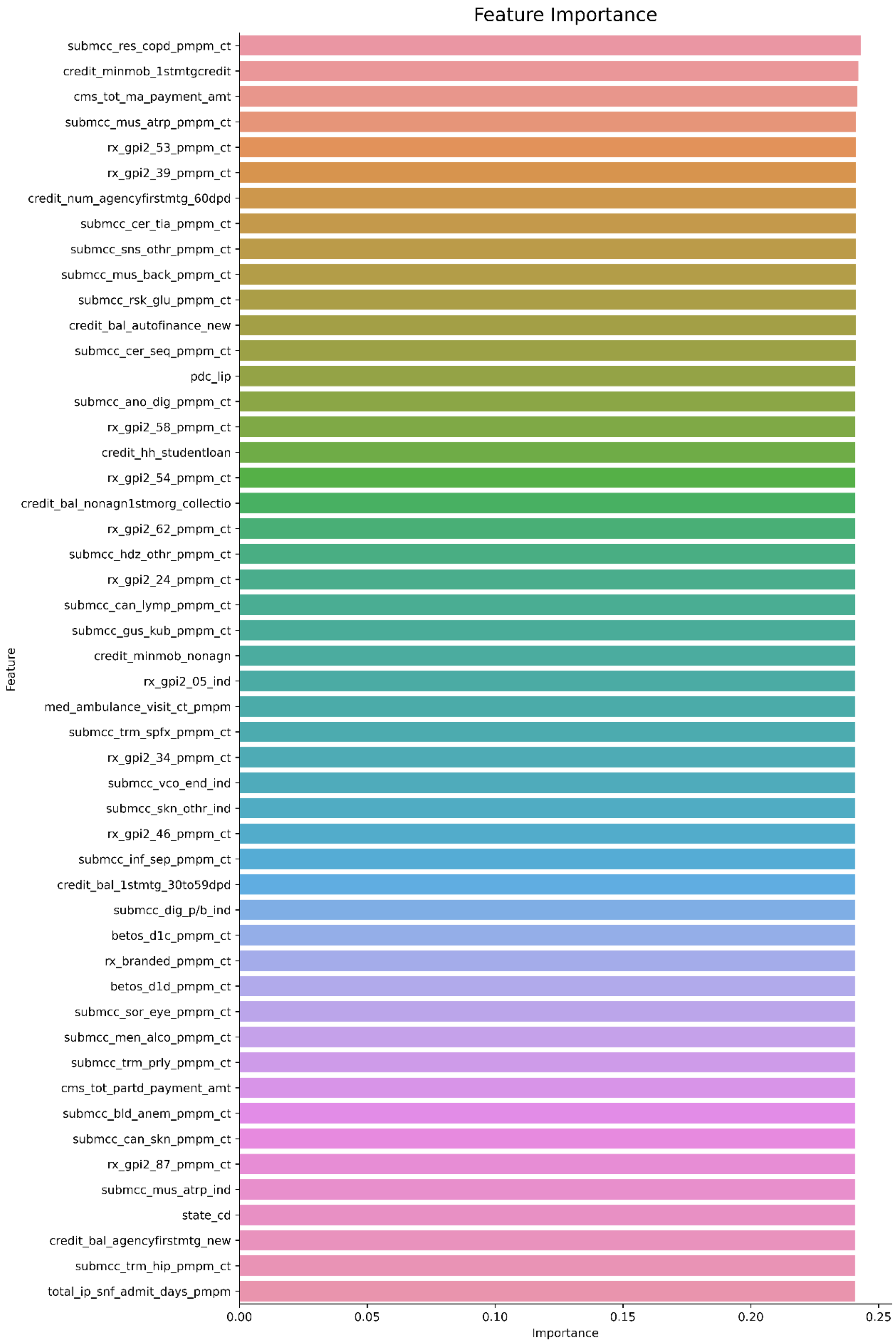


Figure 5: Feature Importance of Top 50 Features

5. Recommendation and Actionability

5.1 Benefits of Releasing Transportation Challenges

5.1.1 Help Members in Need

Transportation challenges can have a significant negative impact on the quality of one's life. The cause of transportation issues may differ from person to person. Overweight people are facing challenges because of their shapes. People encounter a sudden financial problem may not afford taxi fees anymore. Severe diseases also deprive people the chance to live a satisfying life. Releasing transportation challenges is an opportunity to help these people. By providing members in trouble with support they need, Humana can help improve their life quality.

5.1.2 Help Members in Advance

Also, analyzing the problem gives Humana an insight that which members are more likely to have transportation issues. Humana can help members with higher risk in advance, even before they aware of the problem themselves. Also, Humana can send mails or advertisements to members who have similarities with high risk members in certain aspects. For example, reminding overweight people to keep a healthy diet and exercise more is a warm action, establishing a good image as well as reducing the future cost when those members have severe illness.

5.1.3 Build Brand Image

Releasing transportation challenges can not only help members now in trouble but can also take care of potential high-risk members. This will help Humana to establish a brand image that Humana is not only a health insurance company balancing risk for you but also willing to help members to live a better life.

5.2 Results Analysis According to Feature Importance

We analyzed the top 300 features in section 4.2 whose importance is over 0.2 and concluded their common characteristics as follows.

- The most frequently appeared features are those begin with "submcc", which means they all belong to the subcategory of major clinical category. Among all these features, several occurred many times, indicating they are highly related to "transportation_issues". We concluded all the key words as follows.
    - <u>Musculoskeletal disease</u>, referring to those disease that related to muscle or skeleton.
    - <u>Trauma/Fractures</u>, referring to any physical damage to the body and sometimes may take a long time to recover.
    - <u>Congenital malformation</u>, referring to those disease that has a direct relation with the genetic factor and environmental factor in early pregnancy.
    - <u>Malignancy</u>, referring to a malignant mass of tissue in the body.
    - <u>Heart disease</u>, referring to an abnormal organic condition of the heart or of the heart

and the circulation.
  - Cerebrovascular disease, referring to those disease that related to the blood vessels and the blood supply of the brain.

- Features begin with "credit", which mainly related to credit accounts for mortgage, are also frequently appeared. After carefully review, we found the following three points are very important:
  - Any characteristics that related to the first-time mortgage, such as the balance, the number, whether it's agency or non-agency, etc.
  - Days past due (dpd).
  - The credit for Household (HH) accounts.

- Features begin with "rx", especially those with the format of "rx_gpi2_xx_pmpm_ct" occurred nearly 60 times in all top 300 features. These are prescriptions broken out by GPI2 (General Paresis of Insane) category, meaning that GPI is highly related to “transportation_issues".

- Other demographic features which related to our daily life demonstrated high importance too, such as gender, age, state, mobile vehicle ownership, mobile homes, etc.

5.3 Recommendations

As discussed above, features related to medical history, credit data, specific prescription and demography are crucial and worth to be paid attention to. Therefore, we proposed target recommendations based on these categories.

5.3.1 Members with a Specific Medical History

If Medicare members own medical histories for musculoskeletal disease, trauma/fractures, congenital malformation, malignancy, heart disease, cerebrovascular disease, always go to hospitals to see a doctor or prescribe medication like GPI2, they may have a higher tendency to be disposed to transportation challenges since most of the Medicare members are over 50 years old. Thus, necessary resources and service are needed.

- Remind them to conduct regular medical check-ups in case of any relapse due to the difference in temperature between day and night, high-pressure work, a careless diet, etc. These relapses could significantly influence their travel plan.

- Initiate relevant workshops for them to equip them with necessary medical knowledge, such as how to efficiently report medical expense, first-aid knowledge, etc., and also some useful tips for daily transportation, such as how to taxi-hailing apps since most senior members whose age is over 70 years old may lack the expertise of using it.

5.3.2 Members with Credit Issues

Members with credit issues such as a long dpd to repay loan or mortgage, having household accounts credit issue, etc. may have a high tendency to face financial pressure currently or in the future. Thus, necessary interactions and service are needed.

- Organize activities such as one-to-one talk, group talk, encouraging members to speak out their worries with each other. This may serve as a good place for members to exchange their worries, ideas and experience to get through all those hard times.

- Initiate mentor-mentee programs. Hire experts from financial area to serve as mentors to those who currently face and want to solve financial problems. The program may charge some fees depends on the demands.

5.3.3 Members with Specific Demographic Characteristics

Members who are in specific demographic categories such as a senior age, a bad habit, a specific educational level, etc. may have a high tendency to be disposed to transportation challenges. Thus, necessary interactions and resources are needed.

- Set Charted buses for those members who lived together in a rural area where is lack of convenient transportation means like subways.
- Post articles/courses regularly on Humana's official website, involving subjects like health tips, food for health, mental health, etc. and share these articles to members with bad health habits (e.g. smoking).
- Create a forum to talk about health-related topics and invite experts to share their opinions, attracting more people to come by.

5.4 Actionable Steps

1. Pay close attention to members who have medical histories for musculoskeletal disease, trauma, congenital malformation, malignancy, heart disease or cerebrovascular disease. Humana can help them by arranging regular medical check-ups and give them some advice so that they can have a quicker rehabilitation.
2. Workshops on day-to-day transportation tips should be held. For example, elder members may need some knowledge of how to use taxi-hailing apps.
3. Members with credit issues also need special care both mentally and practically. Humana could encourage members to speak out their worries with each other during group talks. In addition, we can also set mentor-mentee programs to help them actually get rid of severe financial problems once and for all.
4. Setting shuttle bus lines among communities in rural areas, especially where there are no metros will help solve transportation issues.
5. At the meantime, Humana could provide its members with platforms such as online forum to offer knowledge about health and with channels via which members can communicate with each other about healthy lifestyle and good habits.

5.5 Future Considerations

1. Setting clear criteria or definition for transportation issue is urgent. The definition of transportation issue now is relatively ambiguous. Since it is self-reported, the criteria for transportation issue may vary from member to member. So, the outcome of the survey may

not be as reliable as it is supposed to be. Also, instead of self-reporting, Humana can assign a group of staff to identify issues related to transportation so that the outcome can be much more consistent and reliable.

2. Supplement with geographic data. It is important to gather and analyze geographic information when predicting transportation challenges. Now, we have 4 related features (rucc_category, state_cd, zip_cd, cnty_cd), but only the first feature does not have large number of missing values or unclear values (e.g. 'other'). Missing values do make it tough for us to locate every Medicare member and discover some valuable insights about where they live, if there are some convenient facilities/means of transportation nearby, is there home far from their workplace, etc. Access to the missing geographic information could help to further improve the accuracy of our model and Humana could thus providing target resources and service for members in different areas.
3. Humana needs to include time-series information to the data. The physical, mental and financial condition of a certain member cannot be constant. The information we currently have is "who and how he is now" that makes him have transportation issues. What is more important is that "what has changed" that leads him to transportation issues.

6. Conclusion

In transforming Humana's longitudinal data which contains 800+ features to cleaned data, we tried several methods and concluded that natural language processing technique could extract the most comprehensive information behind features. And a deep neural network model could predict Medicare members who have a transportation issue with an AUC of .7406. Comparing this model to traditional models such as Random Forest and Extreme Gradient Boosting, and another deep model DeepFM, it is clear that our final model outweighs others because of higher AUC as well as efficient performance. The variable importance of our model offered us interpretable insights that allowed us to formulate recommendations related to sharing information with Humana. We believe that this model's ability to identify and predict members who are struggling with transportation challenges could motivate Humana to better provide members in trouble with support they need, remind those with higher risk in advance, and establish a brand image for Humana.

Appendix

Please check out the complete python code in the following link:
(https://github.com/liyue34673/Humana_2020_case_competition)