**Slide 1 - 1**

Good Evening Vikyath, Hope you had a fascinating day. I am Mohanaditya Karampudi or you can simply call me MO. I am an Applied Data Scientist by training but at the moment I am a graduate student in Business Analytics course at University of South Florida and a TA for master level statistics course. I have nearly 3 years of work experience, the first six months as Software tester and the last two years as Junior Data Scientist. All of this is back in India.

Today we will have convesation about a competition I have participated in my third semester. We were a team of three and was a four week competition. It is the Humana Health care Competition and this was the project we did beyond the coursework requirements.

**Slide 2 - 1**

I will start this presentation by detailing about the problem, next give an overview of the type of features, values data hold. Then I will drive through various cleansing and Visualization plots that reveal descriptive statistics of the features and also provide strong reasons for feature selection. Later on, its about the details we looked into for building and fine tuning the model. I will also share details about best metric choosed to evaluate the performance of the final model and detail the important features, their shapley interpretation and finally present the ways this approach can be improved.

**PSlide 3 - 2**

Before we dive into the data let me first give a high level view of how machine learning works. With the advent of computers it became really easy to automate the monotonous tasks which require simple rules to work.

Consider this simple example, the first four columns are the points of a shape, these contain values 0 and 1 where 0 represents there is no point and 1 represents a point.

If there is just a single one then it is a point and if there are two one then we can draw a line and 3 points then a triangle and finally with 4 points we can draw a quadrilateral.

Now, this is a linear separable data, we can simple hard-code these rules to create a system capable of classifying the future data.

We can also solve this problem by building a Machine Learning model that can automatically find such rules to successfully classify. This is fairly simple data but, out in the real-world it is not easy for humans to come up with the rules for different type of problems.

Even with years of expertise professionals like doctors might make mistakes because the rules they learnt to operate might not be the best solution in that scenario.

With Increase in computation and data storage capacities, it is now the era of building smart algorithms that are capable of mining the data to come up with such rules. One such real world usage is this competition.

**Slide 4 -2**

The humana is a health insurance company and they is interested in understanding the social determinants of health for their members.

Social determinants of health are the conditions in the environments in which people live, work, and play, that affect the quality of-life.

Humana is seeking a “broader view“ of their members to better understand the whole person and to assist them in new ways towards achieving their best health.

Transportation is one of these determinants.

The goal is to identify Medicare members most at risk for a Transportation Challenge and propose solutions for them to overcome this barrier to accessing care.

Now the dependent variable is a binary classification. They derived this feature by asking a simple question to the members and recorded their response. The question is “In the past 12 months, has a lack of reliable transportation kept you from medical appointments, meetings, work or from getting things needed for daily living?” Yes / No

**Slide 5**

It is a one year data that contains information about the members enrolled in the Humana Medicare Advatage and Prescription drug plan

The dependent variable transportation issue is class imbalanced where 85% of the members do not face transportation issue. This percentage is a very good indication of the robust transportation services in this country. But the goal is to concentrate on understanding the factors that effect the 15% of people.

The dataset shared for building solution has nearly 70K rows and 826 columns where each row represents information about a member thus presenting no auto-correlation and columns reveal details about one of the five categories.

As the source of data is genuine, they have augmented it. Hence, the objective is to create a valid process rather than relying on the numbers in the outcome.

There might be a lot of reasons for transportation issue and the goal is find out what factors from the following four categories effect.

1. Medical claim features -
2. Pharmacy claim features
3. Demographics of the members
4. Condition related features

Most of these categories hold two types of features

1. IND - Binary Indicator whether the member availed the service or not
   1. eg. did the person make a specific medical claim or not?
2. PMPM - (Per member per month) How much amount did it cost when compared to the monthly average of all members
   1. eg. If the member requested a specific medical claim,h how much did it cost

**Slide 6**

The goal of this project is to use just the health characteristics of members to predict transportation issue. Sure there might be a lot of other parameters that effect transportatio issue but this is the data we had.

* **Medical claim features** - These provide information about the type of medical procedure, diagnosis tests the person has under gone.

* **Pharmacy Clim features** - These hold information about the type of drugs the members used.

* **Consumer data** - It contains information about the personal attributes of members and also data from the center for medicare and medic aid

* **Condition related features** - These talk about the mental health and also the presence of two more or medical conditions in a patient.

All these categories have many different sub-categories and we will drive through the descriptive statistics and features generated for each of these categories

**Slide 7**

**CCSP**

For Medical claims, there can be 100's of procedures that surgeons do and it is not feasible to save such data. Hence, the health department has come up with CCSP codes that bins the type of procedure into few categories.

The type of medical procedure a member has gone through can aid in predicting the transportation issue..

The data we had has 21 such categories

We have created two new features

* A binary indicator - whether the member used any of the 21 CCS service
* Total number of procedures a member has availed out of 21 CCS services

The plot reveals information about the most frequent procedures undergone by members.

There are 34K members who have gone through birth asphxia procedure and 4200 of them face transportation issue and so on.

Next are the Betos, which hold information about the expenditure incured across categories. For example, the test category contains information whether the member availed dialysis test or blood count test etc. The type of tests conducted on the member provides insights into the medical conditions that may affect chances of facing transportation issues.

There are five main betos categories

* Test - These hold information about the dialysis test or blood count tests est
* Eval - The inormation about regular office visit, or emergency room visit etc

We have created two features:

* yes\_no - did the member use any service in each category
* yes\_no\_sum - total services a member used across all the categories

There are 57K patients who have undergone various tests and 9.6K members suffered from transportation issue

**Slide 8**

The pharmacy claims contained three main type of attributes

prescription days covered - it is the proportion of days of dosage covered by the member. Unfortunately it contained NA values so we discarded it

The features we had are mailed/non-mail - that is if the mails are delivered through mail or did the member pick them up at the nearest store. This can be of help cause if the member requests for mailed way of getting the drugs then maybe the member might face transportation issues

Next are the drugs presribed by the doctor. There will be specific drugs which are not recommended to have while driving and so on. Hence, we felt to include information about drugs in our model building process.There are 95 different drugs that the members have used and this is too many attributes that increased sparsity. So,

* we binned the drugs based on their usage and created 18 new categories after talking with pharma expert
* Also, the count of number of categories of drugs used by members
* The sum of count of PMPM in each category

From the plot we could conclude that cardio and hormone related drugs were mostly used by the members. Out of 54K members who used cardio, almost 8K members had transportation issue

**Slide 9**

The Consumer related features include data about the personal information of the member especially their characteristics. These can provide sufficient information which can help better understand the transportation issue.

The plot to the right is the distribution of age broken by dependent variable.

The interesting thing to note is the mean age of members facing transporation issue is 66 years and those who do not is 71 years. This gives a clue that lower age members have higher probability of facing transportation issue.

The table has the characteristics of the members and the percentage of them facing transportation issue, It is pretty much inline and uniformly distributed.

The next are behavior set of features which has information about the mental health of the members. These set of attribute shed light on overall health of the member and thus contribute to transportation issue.

There are seven categorical features which say whether the member has undergone that specifi type of behavioral issue.

Here we have created two new features

* A binary indicator whether member has faced any of the issue or not
* The numeric feature, the number of behavioral issue a member faced.

From the plot we can conclude that most members face dementia and other anxiety behavioral issues.

**Slide 10**

Multiple Comorbidity is the state in which a member suffers from more than one illness at the same time and this helps understand the complexity faced by the members that effect the overall transportation issue.

There are such categories containing 171 categorical and their respective 171 numerical features and each category contains multiple sub-categories.

For example, one of the 28 categories is Digestive and it holds ind and pmpm of upper-gastro, lower-gastr, liver, pancreas and other gastronomical related information.

Three main set of features created

* Whether the member suffered any of the issue from each category MCC or not
* The count of number of MCC suffered in each category
* The sum of pmpm for each category

The most frequent issue faced by members are Musculo-skeletal MCC. Out of 50K who faced this issue 9K members face transportation issues.

**Slide 11**

As the number of features are high and the data has been parted by multiple categories where each category was containing good amount of columns we have first created individual models for each category and select top performing attributes which resulted in a final 138 attributes.

This was not the best approach as there is loss of interaction terms and the rules made by the model are pertinent to just the

sample of columns. But we decided to go ahead due to the limited computing resource and time. **An xgboost algorithm took nearly 5 hours and did not converge. So, parameter tuning was a big challenge in using the whole dataset. Also, I do not prefer to use kitchen sink model to select features.**

Next is the train-test split. We went with a traditional split with 85-15.

As the dependent variable is class imbalanced, there is high chance for model to learn the rules pertinent to just the major class and hence perform poorly. So, we have employed three types of sampling techniques, undersampling, oversampling and SMOTE. The oversampling worked best for us.

Before model building, we have standardized the numerical attributes and label encoded the categorical features.

As this is a supervised classification problem, we have implemented logistic, balanced bagging, random forest, xgboost and light gbm with 5-fold cross-validation and paramteter tuning. The Light-GBM performed best.

These are the final set of parameters for the model.

**Slide 12**

One important way to understand how well the model rules pertain to the real-world is to do testing. While there are many metrics that reveal the performance of the model, for this problem AUC is the best one to look at. As this is class imbalanced one, the accuracy is not a reliable performance metric.

While all the models performed similar, the two models that generalized well are light gbm and logistic regression. I have presented the metrics below and it is remarkable to see an AUC of 0.73 with logistic regression which is a simple model. This is the best we can do with simple model on this dataset.

Then the light GBM has improved the AUC further to 0.78 on train and 0.77 on test. There is slight overfitting but overall it did a great job in generalizing. If we see the other metrics, they are drastically different between train and test, because the test dataset is a class imbalanced data i.e the kind of data we see in the real world.

**Slide 13**

There should always be a lot of emphasis on how was the model able to achieve the result. What features contributed for this outcome. This methodology further improves our reliability on the models prediction.

We have implemented Shapley on a sample of 5000 rows. The plot shows the contribution of most important attributes.

There are two main components in deciphering one is the color which tells the intensity of contribution and the other is direction in which it has impact.

. The high positive contributors are

* CMS part d payment
* CMS disabled indicator
* ccsp superficia injury
* betos other ambulance

The high negative contributors are

* age
* ccsp hypoxia
* betos evaluation indicator

**Slide 14**

The first plot shows the shapley contribution for ambulance service on transporation issue. The x-axis contains the pmpm value of ambulance service. The y-axis shows the probability of facing transportation issue. We can observe there is a positive relationship where transportation issue increases with increase in pmpm value. This can be further grilled out to the specific threshold. If the ambulanace pmpm for a member is greater than 1 then the probaility of that member facing transportation issues will be increase.

Next is the age.

it is further evident in this plot, as seen from the descriptive statistics, that the increase in age reduces the chances of facing transportation issues especially the transportation issue goes down to members having age greater than 65. This might be due to the best medicare and old home services offered.

There is an interesting plot on the interaction of the diability and cms part b type of payment. If a member is disabled then the increase in predictability of transportation is almost zero irrespective of the increase in payment\_rate\_b (as most of the red points are on the zero line), but if the member is not disabled and the cms\_risk\_adj\_payment\_rate\_b\_amt is greater than 500 then the predictability increases.

**Conclusions:**

I really cannot do justice in trying to cover all the details of the project in this limited time. So, I have picked the most importatnt categories that are highly connected to the dependent variable. There are actually eight categories. The aim for this presentation is to show the complexities we faced and the approach used to deal. This is

LightGBM uses a novel technique of Gradient-based One-Side Sampling (GOSS) to filter out the data instances for finding a split value while XGBoost uses pre-sorted algorithm & Histogram-based algorithm for computing the best split.

**Questions to the Data Science Presentation**

* It is so great to see firm open-sourcing the core applications used. The Kedro, casualnex, performanceAI and studio. I guess they must be pretty easy to use and get hands-on. If you have worked in any of these projects, can I know the way it works? Will the team first build project and then open-source it or will someone comeup with the idea of creating a project and take the initaition?

* I might be numerous projects that you have worked or headed, I am interested to know more about the industry and clients openness to use deep learning models rather than machine learning. Sure these two are quite differenct when compared to the applications and the spaces where they are used. Can I know the overall outlook of the firm towards the machine learning and deep learning especially how open is it to embrace new technologies and improvements?

* It is quite fascinating to see the firm's interest in joining NIPS conference every year. If you have participated, can I know the process for participation. Will the employees register and share their latest work or will the company chose the most appealing advances made in the firm?