# HUMANA-MAYS ANALYTICS

## Round 1 Submission


### Abstract

This document reports the problem statement, analytical approach for problem solving and results from quantitative analyses performed to determine if a patient is going to have an AMI



Team Name - Novel Analytics

# Analytical Approach

Problem Statement: Given a population of Medicare Advantage members continuously enrolled in the previous year and with no indication of an AMI, can you predict who is most likely to have an AMI in the next 3 months?

Problem Definition: All the information (all the columns/variables) that was available for Medicare Advantage members was split into three buckets – **Patient Attributes, Medical History and Insurance Coverage.** Using research, hypotheses were formulated that stated the impact of factors from the three buckets on chances of a person likely to have an AMI in the next 3 months.

**Summary :** We cleaned the data and removed observations where there were missing values. We also checked the distribution to understand what constitutes as abnormal values for a variable and if there is any outlier removal required to perform for those cases. Next, we did basic EDAs and formulated/validated hypothesis.

As the problem required for us to predict customer who are at potential risk of heart loss, we looked at various techniques like (Classification) gradient boosting, logistic regression but after looking at preliminary results decided to go ahead with Random forest.

We divided our dataset in test and train (randomly keeping AMI flag 1s same) and took a part of the data to train the model. The results were encouraging and later we decided to optimize the threshold to give equal weightage to 0s and 1s in the predictor(AMI flag).

This is what the result we are getting on Train and Test data.

| | | Final Model | |
|---|---|---|---|
| | | Actual | |
| Predicted | | 0 | 1 |
| | 0 | 49131 | 751 |
| | 1 | 27682 | 1351 |

| | |
|---|---|
| Correct Predictions | 50482 |
| True positives %(TPR) | 64% |
| True Negative % (TNR) | 64% |
| Positive Predicted Values | 5% |
| Negative Predicted Values | 98% |

| Predicte | Actual | |
|---|---|---|
| | 0 | 1 |
| 0 | 8509 | 0 |
| 1 | 26 | 233 |

| | |
|---|---|
| Correct Predictions | 8742 |
| True positives %(TPR) | 100% |
| True Negative % (TNR) | 100% |
| Positive Predicted Values | 90% |
| Negative Predicted Values | 100% |

Data Cleaning: Univariate analysis was performed on all the variables to check if the variables to check their distributions, missing and null values. Appropriate measures were taken to treat the missing values.

Hypothesis Testing: Comparing proportions of members with an existing AMI flag and members with no AMI flag helped in confirming to certain hypotheses

Data Modeling: The cleaned dataset along with the list of variables selected from hypotheses testing were taken as inputs for data modeling. Random Forest classification technique was used to arrive at the flags for members who are most likely to have an AMI

## Problem Definition

Below is the list of all the hypotheses that were created for the three buckets.

| Category | Hypothesis Description | Variables Impacted |
|---|---|---|
| Patient Attributes | Risk of Heart Disease increases with age | AGE |
| Patient Attributes | Men have higher risk of heart disease than females till age of 60, After that both have same | SEX_CD |
| Patient Attributes | Higher BMI will have higher chances of heart attack | EST_BMI_DECILE |
| Patient Attributes | Tends to Struggle with Medical Language Decile; Lower the decile lower chances of heart attack | DECILE_STRUGGLE_MED_LANG |
| Patient Attributes | Educated people are less likely to have heart attack as they will read the sideeffects on prescriptions if any are there | EDUCATION_LEVEL |

| | | |
|---|---|---|
| Patient Attributes | People with higher income have higher chances of heart attack | Census % Below Poverty Line ,Census % Above Poverty Line ,Z4 Home Value ,Estimated Net Worth ,Estimated Income |
| Patient Attributes | People living in metros have hisher risk of heart attack | POPULATION_DENSITY_CENTILE_US,POPULATION_DENSITY_CENTILE_ST |
| | | |
| Medical History | Higher the Medical Risk and Rx risk higher the chances of getting an heart attack | RECON_MA_RISK_SCORE_NBR;RECON_RX_RISK_SCORE_NBR |
| Medical History | If patient has diabetes then there is higher risk of heart attack | DIABETES |
| Medical History | Type I diabetes is more dangerous and increases risk of heart attack | Diab_Type |
| Medical History | If patient has renal disease then there is higher risk of heart attack | ESRD_IND |
| Medical History | Higher the hospital visits higher the risk of heart attack | CON_VISIT_XX_QYY |
| Medical History | Higher the medical/Rx score higher the chances of heart attack | |
| Medical History | Certain Prescription drug classes increase heart attack risk more than others-Anti-depressants | RX_THER_AA_YR2017 |
| | | |
| Insurance Coverage | Customers enrolled for medicare/medicaid have higher risk of heart attack | |
| Insurance Coverage | Customers taken both Prescription and medical insurance plan has higher risk of heart attack | |
| Insurance Coverage | Coverage type also increases chances of heart attack | |
| Insurance Coverage | Members enrolled in CDC are less likely to get diabetes or Hearth attack | CDC |

## Data Cleaning:

Final data has 87,683 rows which we got after removing observations from original dataset (100k records)

Key Observations:

1. Imputations were done for the following variables: <enter variables>

2. About 12,000 rows which had no values in almost all the columns were removed

3. The final dataset consists of 2335 out of 2726 members with AMI_FLAG = 1

## Hypothesis Testing:

Below are some of the hypotheses that were tested before selecting variables and running the model.

1. Age Groups vs Average % of AMI occurrence:

| AMI Flag | Average of AGE |
|---|---|
| 0 | 72.70159549 |
| 1 | 75.22707263 |

| AMI Flag/ Age Groups | 40-49 | 50-59 | 60-69 | 70-79 | 80-89 | 90-95 |
|---|---|---|---|---|---|---|
| 0 | 99% | 98% | 98% | 97% | 96% | 94% |
| 1 | 1% | 2% | 2% | 3% | 4% | 6% |

Although the average age is same is for both affected (by AMI) and not affected; when we check across age the incident rate increases with age which suggests that chances of AMI incidence increases with higher range.

2. Gender

| AMI Flag | F | M |
|---|---|---|
| 0 | 55476 | 41771 |
| 1 | 1292 | 1431 |
| Total | 56768 | 43202 |
| **% AMI Flag=1** | 2.3% | 3.3% |

**For cases <60 age**

| AMI Flag | F | M |
|---|---|---|
| 0 | 5090 | 4626 |
| 1 | 78 | 130 |
| Total | 5168 | 4756 |
| % AMI Flag=1 | 1.5% | 2.7% |

Males have higher risk of heart attack for both cases of age groups.

3. BMI

| AMI Flag | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 6295 | 8000 | 10860 | 11609 | 8032 | 11342 | 9030 | 7550 | 8055 | 4707 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 204 | 281 | 329 | 342 | 224 | 300 | 261 | 167 | 153 | 80 |
| %AMI Flag=1 | 3.1% | 3.4% | 2.9% | 2.9% | 2.7% | 2.6% | 2.8% | 2.2% | 1.9% | 1.7% |

Higher BMI increases chances of heart attack.

4. Medical Language Decile

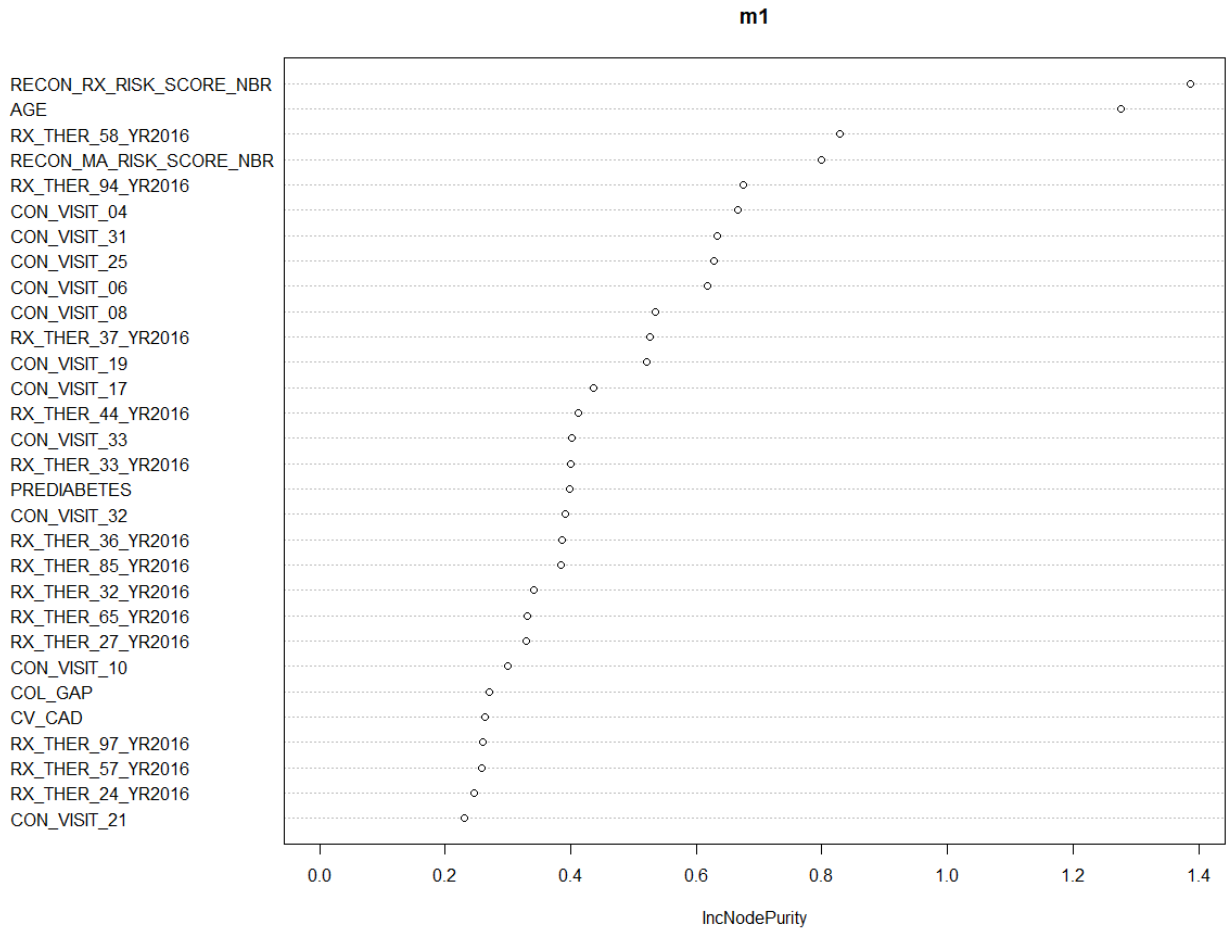
| AMI Flag | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 3257 | 3556 | 4067 | 4885 | 5613 | 6435 | 7807 | 9193 | 11733 | 16244 |
| 1 | 70 | 76 | 74 | 114 | 142 | 165 | 204 | 242 | 346 | 562 |
| | 2.1% | 2.1% | 1.8% | 2.3% | 2.5% | 2.6% | 2.6% | 2.6% | 2.9% | 3.5% |

There is not much evidence to say that people having difficulty in understanding of medical language faces higher risk of heart attack.

# Data Modeling

**Variables that came out significant**

Running the Random Forest classification model gave us the following variables in their order of importance towards the AMI occurrence.



Inferences about the variables->

1) **RECON_RX_RISK_SCORE_NBR** –(Rx Risk Score calculated by CMS) This came out significant and tells what is the risk of the patients starting on prescription drugs. Drugs also result in higher chances of Heart attack
2) **Age**- It came out positive as with higher age the chances of heart attack increases
3) **Rx_THER_58_YR2016** which is the Rx for antidepressant came out significant. Since they are known to increase blood pressure resulting in heart attack
4) **RECON_MA_RISK_SCORE_NBR -**Medical Risk Score calculated by CMS; This came out significant as this is proxy for the patients who may be at risk and will likely be getting a medical treatment including heart surgery
5) **RX_THER_94_YR2016-**DIAGNOSTIC PRODUCTS- As per our understanding this could be capturing patients who have asked to buy some diagnostic product-Heart rate monitor etc which suggests early symptoms of heart attack

**Other conditions in the order of Importance-**

| **Condition Code** | Conditions |
|---|---|
| **04** | OTHER CIRCULATORY |
| **31** | RISK BEHAVIORS |
| **25** | DIGESTIVE |
| **06** | MALIGNANT NEOPLASMS |
| **08** | DIABETES |
| **19** | SENSE ORGANS |
| **17** | CEREBROVASCULAR |
| **33** | V-CODES |
| **32** | SIGNS AND SYMPTOMS |
| **10** | GENITO-URINARY SYSTEM |
| **21** | MUSCULOSKELETAL AND CONNECTIVE TISSUE |
| **24** | RESPIRATORY |
| **26** | DISEASES OF SKIN AND SUBCUTANEOUS TISSUE |
| **05** | DISEASES OF BLOOD AND BLOOD-FORMING ORGANS |
| **23** | INFECTIONS |
| **09** | ENDOCRINE |
| **01** | CORONARY ARTERY DISEASE |
| **02** | CONGESTIVE HEART FAILURE |

6)

| Variables | IncNodePurity |
|---|---|
| RECON_RX_RISK_SCORE_NBR | 1.387156387 |
| AGE | 1.275650543 |
| RX_THER_58_YR2016 | 0.828211307 |
| RECON_MA_RISK_SCORE_NBR | 0.798822619 |
| RX_THER_94_YR2016 | 0.674285492 |
| CON_VISIT_04 | 0.666630063 |
| CON_VISIT_31 | 0.632774481 |
| CON_VISIT_25 | 0.628225748 |
| CON_VISIT_06 | 0.6170933 |
| CON_VISIT_08 | 0.534064689 |
| RX_THER_37_YR2016 | 0.525810601 |

| | |
|---|---|
| CON_VISIT_19 | 0.521913634 |
| CON_VISIT_17 | 0.436847059 |
| RX_THER_44_YR2016 | 0.412851309 |
| CON_VISIT_33 | 0.401067854 |
| RX_THER_33_YR2016 | 0.399424224 |
| PREDIABETES | 0.398748498 |
| CON_VISIT_32 | 0.391487936 |
| RX_THER_36_YR2016 | 0.386682553 |
| RX_THER_85_YR2016 | 0.384296328 |
| RX_THER_32_YR2016 | 0.340967153 |
| RX_THER_65_YR2016 | 0.330373921 |
| RX_THER_27_YR2016 | 0.32909819 |
| CON_VISIT_10 | 0.29920717 |
| COL_GAP | 0.270083774 |
| CV_CAD | 0.262896618 |
| RX_THER_97_YR2016 | 0.259849817 |
| RX_THER_57_YR2016 | 0.258355452 |
| RX_THER_24_YR2016 | 0.246676504 |
| CON_VISIT_21 | 0.230036344 |
| RX_THER_42_YR2016 | 0.220687866 |
| RX_THER_49_YR2016 | 0.21681947 |
| HYPERTENSION | 0.214619408 |
| CON_VISIT_24 | 0.212774825 |
| RX_THER_66_YR2016 | 0.210196442 |
| RX_THER_39_YR2016 | 0.206739696 |
| SEX_CD | 0.188112225 |
| RX_THER_12_YR2016 | 0.179272446 |
| RX_THER_72_YR2016 | 0.179220763 |
| RX_THER_34_YR2016 | 0.168041981 |
| RX_THER_30_YR2016 | 0.164063669 |
| RX_THER_16_YR2016 | 0.150248211 |
| RES_COPD | 0.148487509 |
| CON_VISIT_26 | 0.138500197 |
| RX_THER_17_YR2016 | 0.134593676 |
| Diab_Complications | 0.11097297 |
| RX_THER_22_YR2016 | 0.109361394 |
| CON_VISIT_05 | 0.094037284 |
| CON_VISIT_23 | 0.091539596 |
| CON_VISIT_09 | 0.08792881 |
| RX_THER_82_YR2016 | 0.066762511 |
| CV_CHF | 0.060439848 |
| CV_HDZ | 0.055114416 |
| RX_THER_52_YR2016 | 0.050983881 |
| RX_THER_02_YR2016 | 0.047060563 |
| CON_VISIT_01 | 0.039688251 |
| ARTH | 0.032490141 |
| CDC | 0.029744317 |

| | |
|---|---|
| RX_THER_11_YR2016 | 0.028171722 |
| CDC_HBAPOOR_GAP | 0.027829746 |
| CDC_EYE_GAP | 0.025371934 |
| RX_THER_03_YR2016 | 0.02229694 |
| CON_VISIT_02 | 0.007333333 |
| OSTEO | 0.001777778 |
| RX_THER_21_YR2016 | 0.001714286 |
| RX_THER_60_YR2016 | 0.001666667 |

## Model Technique

After trying Gradient Boosting classification and logistic regression, we final went ahead with Random Forest as this would be able to capture variation in the data and would be much faster than GBM.

## Model validation

## Final Results

**We took 10% of the Cleaned data to train the data due to resource constraint and following is the result we are getting on train data**

| | | Actual | |
|---|---|---|---|
| Predicte | | 0 | 1 |
| | 0 | 8509 | 0 |
| | 1 | 26 | 233 |

| | |
|---|---|
| Correct Predictions | 8742 |
| True positives %(TPR) | 100% |
| True Negative % (TNR) | 100% |
| Positive Predicted Values | 90% |
| Negative Predicted Values | 100% |

**When we use the same model and optimize threshold to give equal weightage to 1s and 0s, we get this result. The results give accuracy of 64% but can capture both 0s and 1s with reasonable accuracy unlike previous models. This will help us identify the customers at risk of heart attack and shape our strategy accordingly.**

| | | Final Model Results( Test Data) | |
|---|---|---|---|
| | | Actual | |
| Predicted | | 0 | 1 |
| | 0 | 49,131 | 751 |
| | 1 | 27,682 | 1,351 |

| | |
|---|---|
| Correct Predictions | 50,482 |
| True positives %(TPR) | 64% |
| True Negative % (TNR) | 64% |
| Positive Predicted Values | 5% |
| Negative Predicted Values | 98% |

**These results can be further improved by taking larger sample as train data which we couldn't do because of RAM constraints and running more iterations to get right set of variables.**