OUD Progression – Building a Predictive Model

# Sample Selection:

Using an external MarketScan Medicaid dataset for services between 2013 and 2016, we selected enrollees with the following requirements:

* t
* Mustboth medical and pharmacy coverage .
* Must not have a diagnosis of Cancer or within Hospice care.

Among these members, those with at least one prescription for opioid drugs in 2014 were further selected as having opioid experience that put them in the overall opioid use at-risk pool. Within this group, the target group for prediction of opioid use dependence was constrained to be those having had a “first” prescription during 2014; that is, having a clean period prior to first prescription in 2014 of at least 6 months.

For this predictive model population, the target outcome being predicted is Opioid Use Disorder as defined by a well-recognized list of diagnosis codes (in both icd9 and icd10) for Opioid User Disorder as well as by prescriptions for opioid treatment drugs, specifically naloxone, buprenorphine or methadone. Members with either of these outcome markers appearing within the measurement year were excluded from the prediction model but counted as members having opioid use disorder within the measurement year. The prediction period is anytime within the year following the measurement year (i.e., anytime in 2015 for the modeled population).

# Training/Testing Set Selection:

Starting from the pre-selected enrollees mentioned above in the ‘Sample Selection’ section, we queried all relevant variables from MarketScan Medicaid data of year 2014, 2015 and 2016 based on their enrollment ID (enrolid).

The variables, including both native and engineered, are listed below by segment:

|  |  |
| --- | --- |
| **Eligibility: Members** | |
| Dataset: elig13\_16\_1 | |
| Variable | Notes |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of 1.1.14 |
| sex | sex as of 1.1.16 |
| recipzip | 3-digit zip as of 1.1.16 |
| stdrace | race as of 1.1.16 |
| msa | msa as of 1.1.16 |
| recipcty | county as of 1.1.16 |
| recipgeoloc | state as of 1.1.16 |
| rural | rural as of 1.1.16 |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Medical (Inpatient + Outpatient): Claims** | |
| Datasets: med14, med5, med16 | |
| Variable | Notes |
| adv | Diagnosis of Adverse Affects due to Opioids |
| cancer | Diagnosis of Cancer (see subcategories used) |
| death | Discharge Status of Expired |
| hospice | Procedure or Revenue Code for Hospice |
| hx\_neo | Diagnosis of History of Malignant Neoplasm (one of the cancer criteria) |
| ip\_op | Inpatient/Outpatient Flag: IP/OP |
| mal\_neo | Diagnosis of Malignant Neoplasm (one of the cancer criteria) |
| mh | Diagnosis of Mental Health Disorder |
| morphine | Procedure Code for Morphine Injection (Rx for Morphine is on the drug claims) |
| nas | Diagnosis of Neonatal Abstinence Syndrome in Infant (Neonatal Withdrawal) |
| nas\_mom | Diagnosis of Neonatal Abstinence Syndrome in Mother (Newborn Affected by Mother's Drug Use) |
| od | Overdose due to Opioids (Poisoning, Accidental Poisoning) |
| oth\_neo | Diagnosis of Other Malignant Neoplasms (one of the cancer criteria) |
| oud | Opioid Use Disorder (Dependence, Use, Abuse) |
| oudr | Opioid Use Disorder in Remission |
| sud | Substance Use Disorder (non-opioid) |
| treat | Substance Abuse Treatment (Pharmacotherapy, Med Mgmt) Opioid/Other Drugs Proc Code |
| treat\_rx | Injection of Naloxone or Buprenorphine to Treat Opioid Addiction |
| ed | ED visit via CPT/revenue code |
| pcp | Primay Care Provider (stdprov = MD, Osteopath, Int Med, Multi-Spec, Fam Pract, Geriatric, Ped, Ped Spec, NP, PA) |
| new\_user\_dt | First opioid Rx date in 2014 (no Rx prior 180d); excl removed (hospice/cancer); else missing |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of svcdate |
| sex | sex as of svcdate |
| svcdate | First date of service |
| tsvcdat | Last date of service |
| stdprov | Provider type |
| stdrace | race as of svcdate |
| rural | rural as of 1.1.16 |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Medical (Inpatient + Outpatient): Members** | |
| Datasets: med14, med15, med16 | |
| Variable | Notes |
| adv | 1+ Diagnosis of Adverse Affects due to Opioids during year |
| cancer | 1+ Diagnosis of Cancer during year |
| death | 1+ Discharge Status of Expired during year |
| hospice | 1+ Procedure or Revenue Code for Hospice during year |
| mh | 1+ Diagnosis of Mental Health Disorder during year |
| morphine | 1+ Procedure Code for Morphine Injection during year |
| nas | 1+ Diagnosis of Neonatal Abstinence Syndrome in Infant (Neonatal Withdrawal) during year |
| nas\_mom | 1+ Dx of Neonatal Abs Syn in Mom (Newborn Affected by Mom's Drug Use) during year |
| od | 1+ Overdose due to Opioids (Poisoning, Accidental Poisoning) during year |
| oud | 1+ Opioid Use Disorder (Dependence, Use, Abuse) during year |
| oudr | 1+ Opioid Use Disorder in Remission during year |
| sud | 1+ Substance Use Disorder (non-opioid) during year |
| treat | 1+ Subst Abuse Trmt (Pharmacotherapy, Med Mgmt) Opioid/Other Drugs Proc during year |
| treat\_rx | 1+ Injection of Naloxone or Buprenorphine to Treat Opioid Addiction during year |
| ed | 1+ ED visit during year |
| pcp | 1+ visit to Primary Care Provider during year |
| new\_user\_dt | First opioid Rx date in 2014 (no Rx prior 180d); excl removed (hospice/cancer); else missing |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of svcdate |
| sex | sex as of last claim |
| stdrace | race as of 1.1.16 |
| rural | rural as of 1.1.16 |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Admission: Admissions** | |
| Datasets: adm14, adm15, adm16 | |
| Variable | Notes |
| treat | ICD10 procedure for therapy for substance (opioid) abuse |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of svcdate |
| sex | sex as of svcdate |
| svcdate | First date of service |
| admdate | admission date |
| disdate | discharge date |
| stdrace | race as of svcdate |
| rural | rural as of svcdate |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Admission: Members** | |
| Datasets: adm14\_1, adm15\_1, adm16\_1 | |
| Variable | Notes |
| treat1 | 1+ procedures for substance (opioid) abuse during year |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of last admission |
| sex | sex as of last admission |
| stdrace | race as of 1.1.16 |
| rural | rural as of 1.1.16 |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Pharmacy Recodes: Claims** | |
| Datasets: (rx13 only for checking clean period), rx14, rx15, rx16 | |
| Variable | Notes |
| antidepressant | Antidepressant Prescription based on generid = 1; else = 0 |
| antipsychotic | Antipsychotic Prescription based on generid = 1; else = 0 |
| bad\_pot | Bad Potentiator based on generid = 1; else = 0 |
| barbiturate | Barbiturate/Sedative prescription based on generid = 1; else = 0 |
| benzodiazepine | Benzodiazepine Prescription based on generid = 1; else = 0 |
| buprenorphine | Buprenorphine Prescription based on generid = 1; else = 0 |
| fentanyl | Fentanyl Prescription based on generid = 1; else = 0 |
| good\_pot | Good Potentiator Prescription based on generid = 1; else = 0 |
| gt7d | >7 Days Supply = 1; else = 0 |
| le7d | <=7 Days Supply = 1; else = 0 |
| long\_acting | Long Acting Opioid Rx based on generid = 1; else = 0 |
| methadone | Methadone Rx based on generid = 1; else = 0 |
| naloxone | Naloxone Rx based on generid = 1; else = 0 |
| naltrexone | Naloxone Rx based on generid = 1; else = 0 |
| nat\_opioid | Natural Opioid Rx based on generid = 1; else = 0 |
| op\_lax | Opioid Laxative Rx based on generid = 1; else = 0 |
| opioid | Opioid Prescription based on generid = 1; else 0 |
| semi\_opioid | Semi-Synthetic Opioid Rx based on generid = 1; else = 0 |
| short\_acting | Short Acting Opioid Rx based on generid = 1; else = 0 |
| syn\_opioid | Synthetic Opioid Rx based on generid = 1; else = 0 |
| muscle | Muscle Relaxant Rx based on generid =1; else = 0 |
| cns\_dep | CNS Depressant Rx based on generid = 1; else = 0 |
| new\_user\_dt | First opioid Rx date in 2014 (no Rx prior 180d); excl removed (hospice/cancer); else missing |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of svcdate |
| sex | sex as of svcdate |
| svcdate | Date Rx filled |
| ord\_prov | Provider NPI who ordered Rx |
| pharmid | Pharmacy NPI that filled Rx |
| daysupp | # days supply of Rx |
| stdrace | race of recipient as of 1.1.16 |
| rural | rural status of recipient as of 1.1.16 |
| enrmon13 | # of months of enrollment in 2013; 11 or 12 |
| enrmon14 | # of months of enrollment in 2014; 11 or 12 |
| enrmon15 | # of months of enrollment in 2015; 11 or 12 |
| enrmon16 | # of months of enrollment in 2016; 11 or 12 |

|  |  |
| --- | --- |
| **Pharmacy Recodes: 1 record per Member** | |
| Datasets: (rx13\_1 only for checking clean period), rx14\_1, rx15\_1, rx16\_1 | |
| Variable | Notes |
| antidepressant1 | At least 1 antidepressant Rx during year |
| antipsychotic1 |  |
| bad\_pot1 | At least 1 bad potentiator Rx during year |
| barbiturate1 | At least 1 barbiturate/sedative Rx during year |
| benzodiazepine1 | At least 1 benzodiazepine Rx during year |
| buprenorphine1 | At least 1 buprenorphine Rx during year |
| fentanyl1 | At least 1 fentanyl Rx during year |
| good\_pot1 | At least 1 good potentiator Rx during year |
| long\_acting1 | At least 1 long acting opioid Rx during year |
| methadone1 | At least 1 methadone Rx during year |
| naloxone1 | At least 1 naloxone Rx during year |
| naltrexone1 | At least 1 naltrexone Rx during year |
| nat\_opioid1 | At least 1 natural opioid Rx during year |
| op\_lax1 | At least 1 opioid laxative Rx during year |
| opioid1 | At least 1 opioid Rx during year |
| semi\_opioid1 | At least 1 semi-synthetic opioid Rx during year |
| short\_acting1 | At least 1 short acting opioid Rx during year |
| syn\_opioid1 | At least 1 synthetic opioid Rx during year |
| op\_benz | At least 1 opioid & 1 benzodiazepine during year |
| op\_lax\_c | At least 1 opioid & 1 laxative during year |
| op\_dep | At least 1 opioid & 1 antidepressant during year |
| op\_psy | At least 1 opioid & 1 antipsychotic during year |
| muscle1 | At least 1 muscle relaxant Rx during year |
| trinity | At least 1 opioid & 1 benzo & 1 muscle relax during year |
| cns\_dep1 | At least 1 CNS depressant Rx during year |
| opioid\_days | sum of days supply of opioids during year |
| new\_user\_dt | First opioid Rx date in 2014 (no Rx prior 180d); excl removed (hospice/cancer); else missing |
| client | client + enrolid = unique ID |
| enrolid |
| age | age as of last claim |
| sex | sex as of last claim |

# Modeling:

The objective is to predict whether a patient will be Opioid Dependent or not within a prediction year (calendar year) based on the data of that patient for the previous year (measurement year, also set to be a calendar year).

In this model, we are using the admission, medical and drug records for 2014 of patients that have a “clean period” in 2013. The target variable here is Opioid Use Dependence (OUD) diagnosis in year 2015.

# Merging the data:

First, we merge the adm\_14, med\_14, rx\_14 files on ENROLIDs and select the data for patients with clean record in 2013.

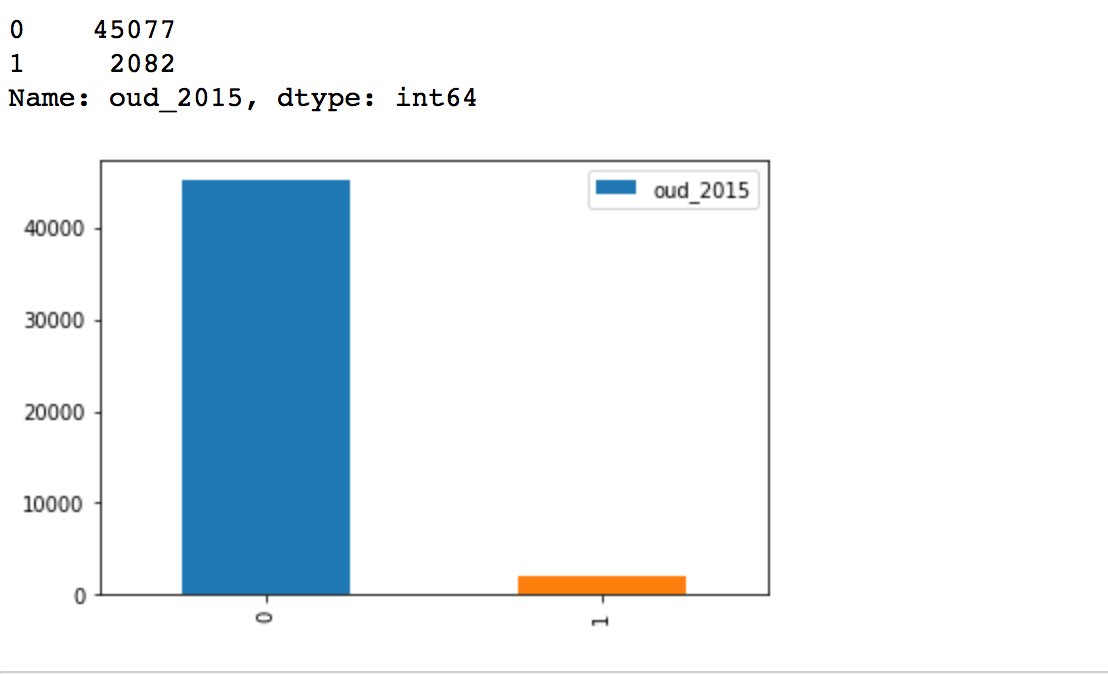
We then add the OUD diagnosis from med\_15 file to the dataset for the common ENROLIDs.

We then check for missing values. We find that we have around 10% missing values, mostly in the ‘STDRACE’ column. We decide to drop all the rows with missing values.

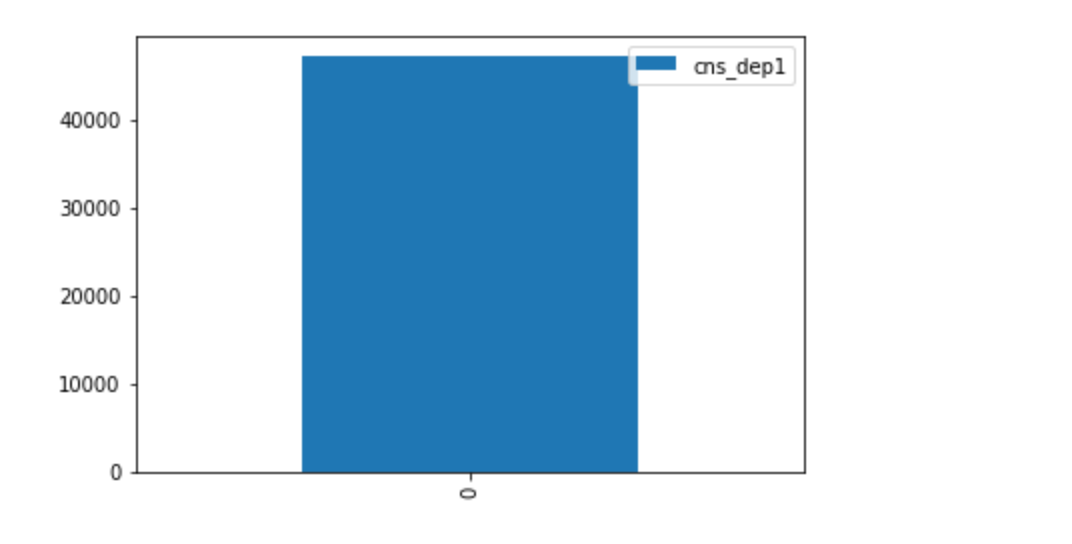
# Exploratory Data Analysis:

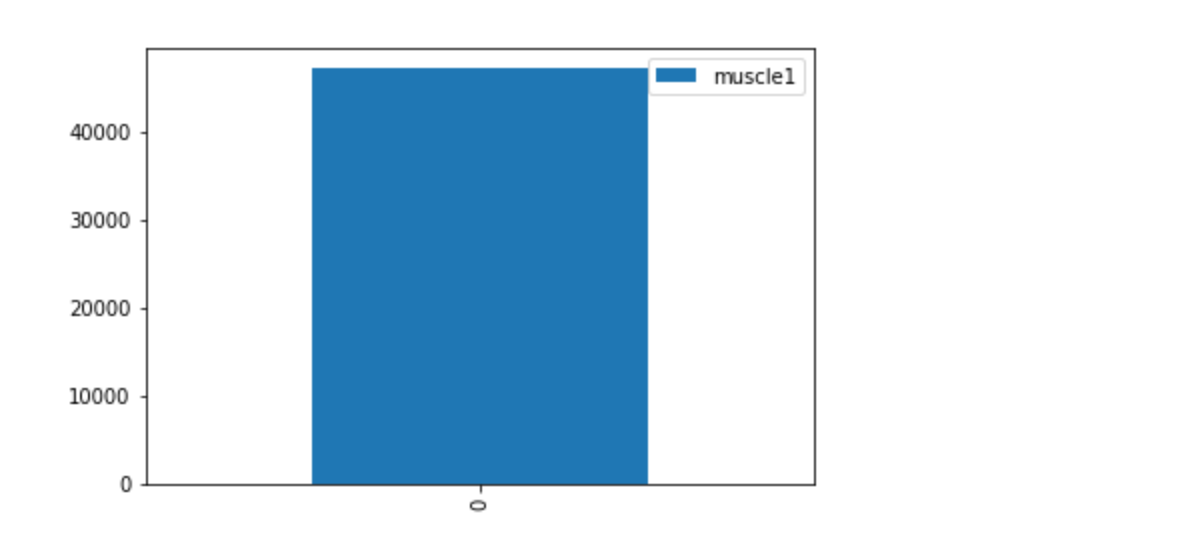
We analyze the data using graphs.

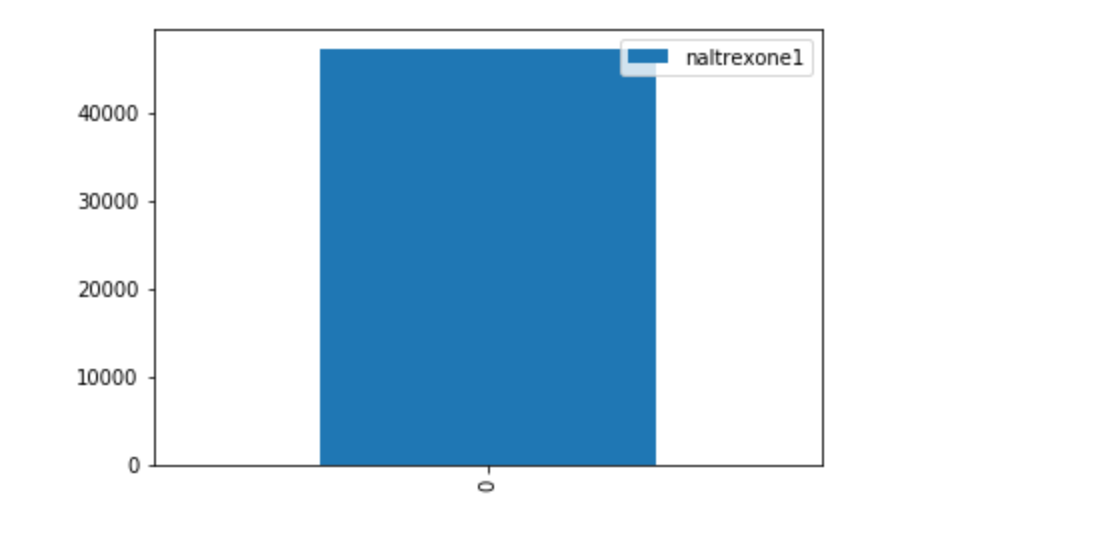
The first thing we observe is that the data is highly imbalanced.



We also observe that some of the columns have only one value, that is, these columns have no variance and hence add no information towards modeling the data. We decide to drop these columns.





****

**To get an estimate of what features are important towards modeling, we estimate feature importance using Random Forest.**

The general idea is to permute the values of each feature and measure how much the permutation decreases the accuracy of the model. For unimportant variables, the permutation should have little to no effect on model accuracy, while permuting important variables should significantly decrease it.

We select the Top 10 variables for modeling the data.

Factors considered in the Opioid Predictive Model (those bolded ends up driving the model):

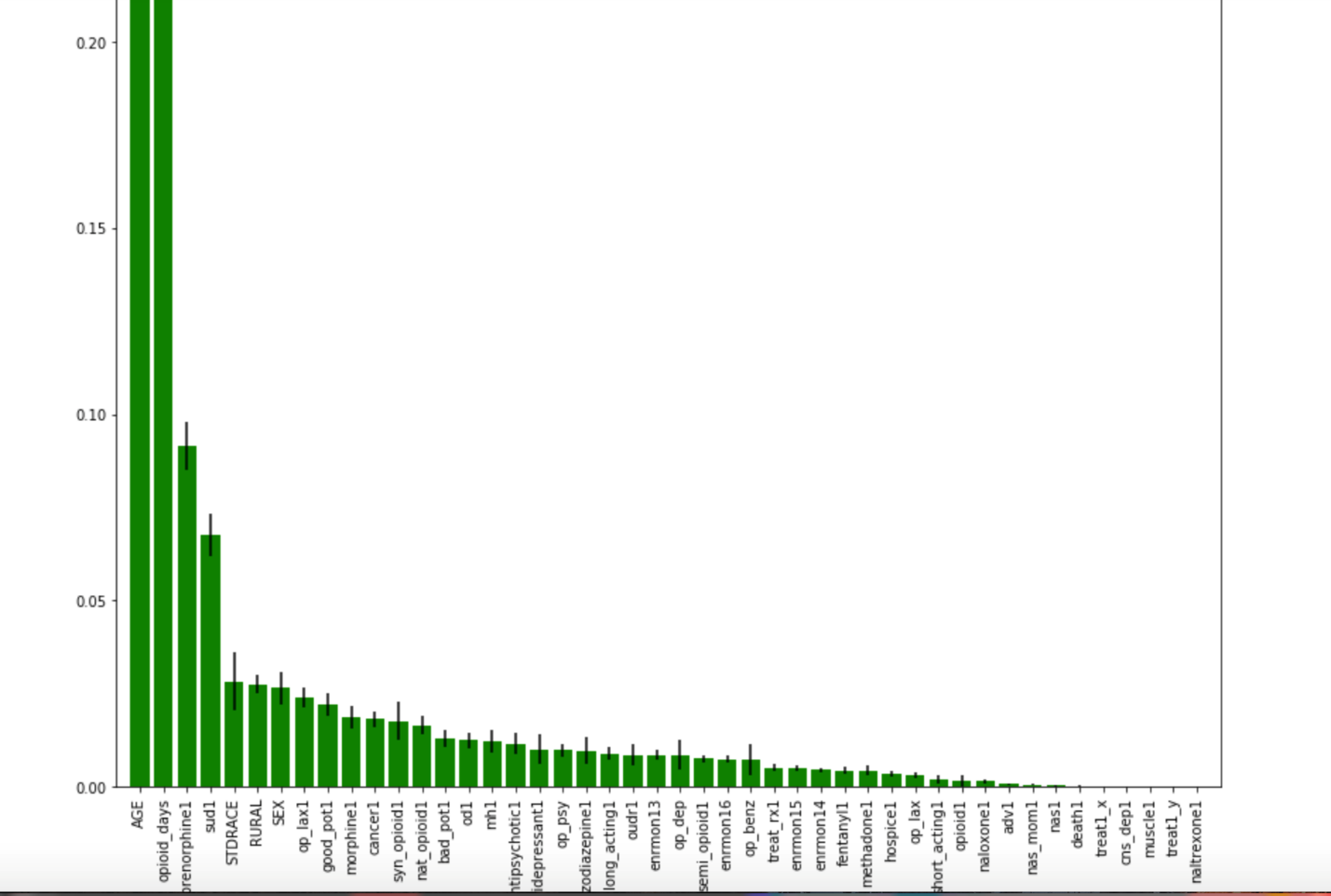
The graph below shows the top-most factors that were considered in the model. This included the factors considered most relevant before making any exclusions to the database. For example, a diagnosis of cancer is high in the list, but this factor was actually treated as an exclusion factor (along with criteria defining patients in hospice care).

The demographic indicators available were rather sparse, but even with this limited palette, 4 among the 10 most important drivers were this category: Age, Race, Rural status, and Sex

The prescription-related characteristics composed the majority of the other model drivers: Total Days Supply for the measurement year being the most influential, and followed by many “drug type” tags associated with the members: opioid-related constipation medication, class of opioid (natural vs synthetic opioid prescriptions), prescriptions for good and bad potentiators, combinations of opioid plus drugs-of-interest (e.g., antidepressants, anti-psychotics, benzodiazapenes)

The diagnosis-related factors were also considered, with a diagnosis of Substance Use Disorder (used for alcohol addiction) being a major driver and a group of diagnoses collectively called mental health being a secondary driver.

Some factors that are strong proxies for OUD in the prediction year and that could be included as drivers were used in one version of the model and explicitly excluded in later versions – these included receiving at least one prescription for buprenorphine (a major factor if included) and the presence of an Overdose crisis event as identified by appropriate diagnosis codes.

****

**Age (from Eligibility file)**

**Opioid days – sum of days supply of opioids during year**

Buprenorphine1 – At least 1 buprenorphine Rx during year

**Sud1 – 1+Diagnosis of Substance Use Disorder during year**

**STDRACE (from eligibility file)**

**RURAL (from eligibility file)**

**SEX (from eligibility file)**

**Op-lax1 - At least 1 opioid laxative Rx during year**

**Good Pot1 – At least 1 good potentiator Rx during year**

**Morphine1 – 1+Procedure Code for Morphine Injection during year**

~~Cancer1 -1+ Diagnosis of Cancer during year~~

**Syn\_opioid1 - At least 1 synthetic opioid Rx during year**

**Nat\_opioid1 – At least 1 natural opioid Rx during year**

**Bad-pot1 - At least 1 bad potentiator Rx during year**

**Od-1 - 1+ Overdose due to Opioids (Poisoning, Accidental Poisoning) during year**

**Mh1 - Diagnosis of Mental Health Disorder during year**

AntiPsychotic1 - At least 1 antipsychotic Rx during year

Antidepressant1 - At least 1 antidepressant Rx during year

Op\_psy - At least 1 opioid & 1 antidepressant Rx during year

Benzodiazapene1 - At least 1 benzodiazepine Rx during year

Long-acting - Long Acting Opioid Rx based on generid = 1; else = 0

Oudr1 - Opioid Use Disorder in Remission during year

Op\_dep - At least 1 opioid & 1 antidepressant Rx during year

Semi\_opioid1 - At least 1 semi-synthetic opioid Rx during year

Op\_benz - At least 1 opioid & 1 benzodiazepine during year

Treat\_rx1- Injection of Naloxone or Buprenorphine to Treat Opioid Addiction

… plus 15 more variables (including the 7+days initial prescription as a factor). This, together with synthetic vs natural, were used in the original IBM research work, but were split out in this modeling construct.

# Data Modeling:

The data that we have is highly imbalanced. There are a few ways to deal with imbalanced data.

While modeling, we can apply class weights to the algorithm which are used to correct the imbalance.

1. Over-sample the minority class to match the majority class.
2. Under-sample the majority class to match the minority class.
3. Generate synthetic samples of the minority class using SMOTE.

We will explore and compare results using these methods.

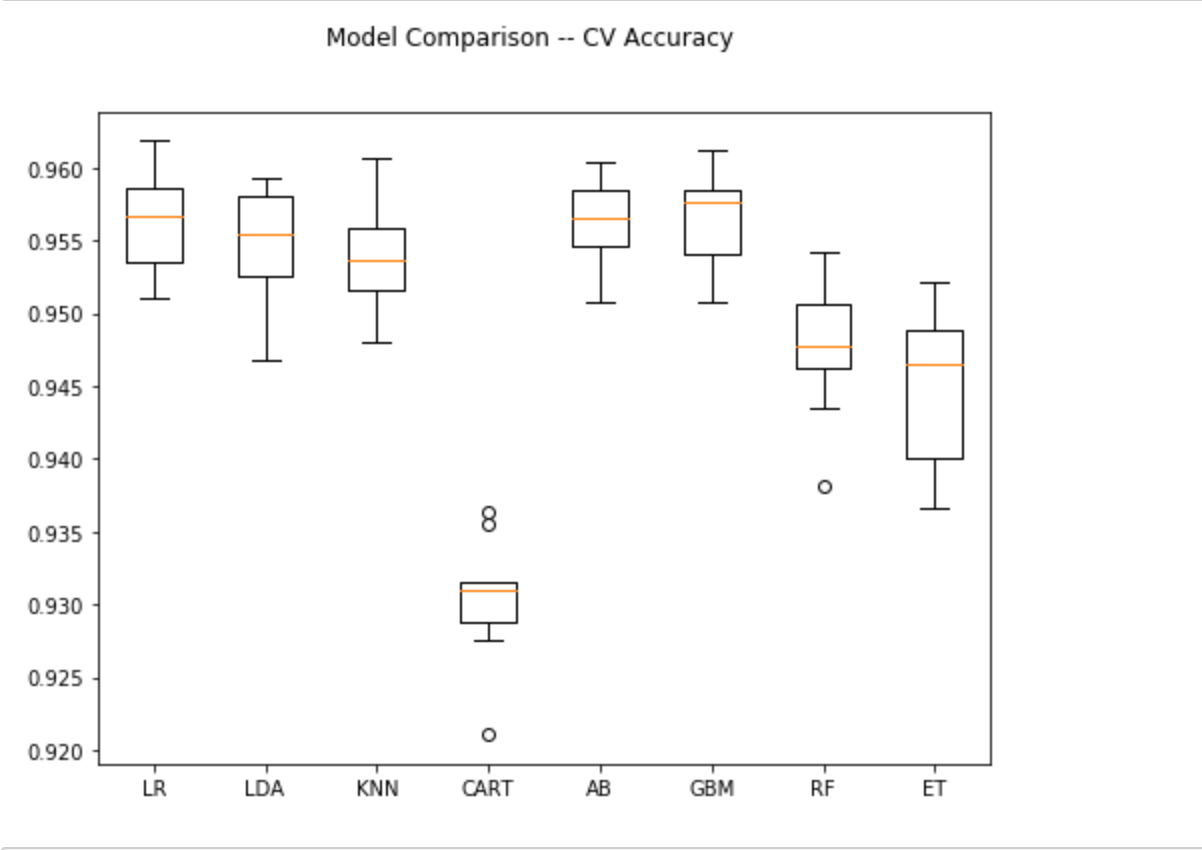
# Using Class Weights:

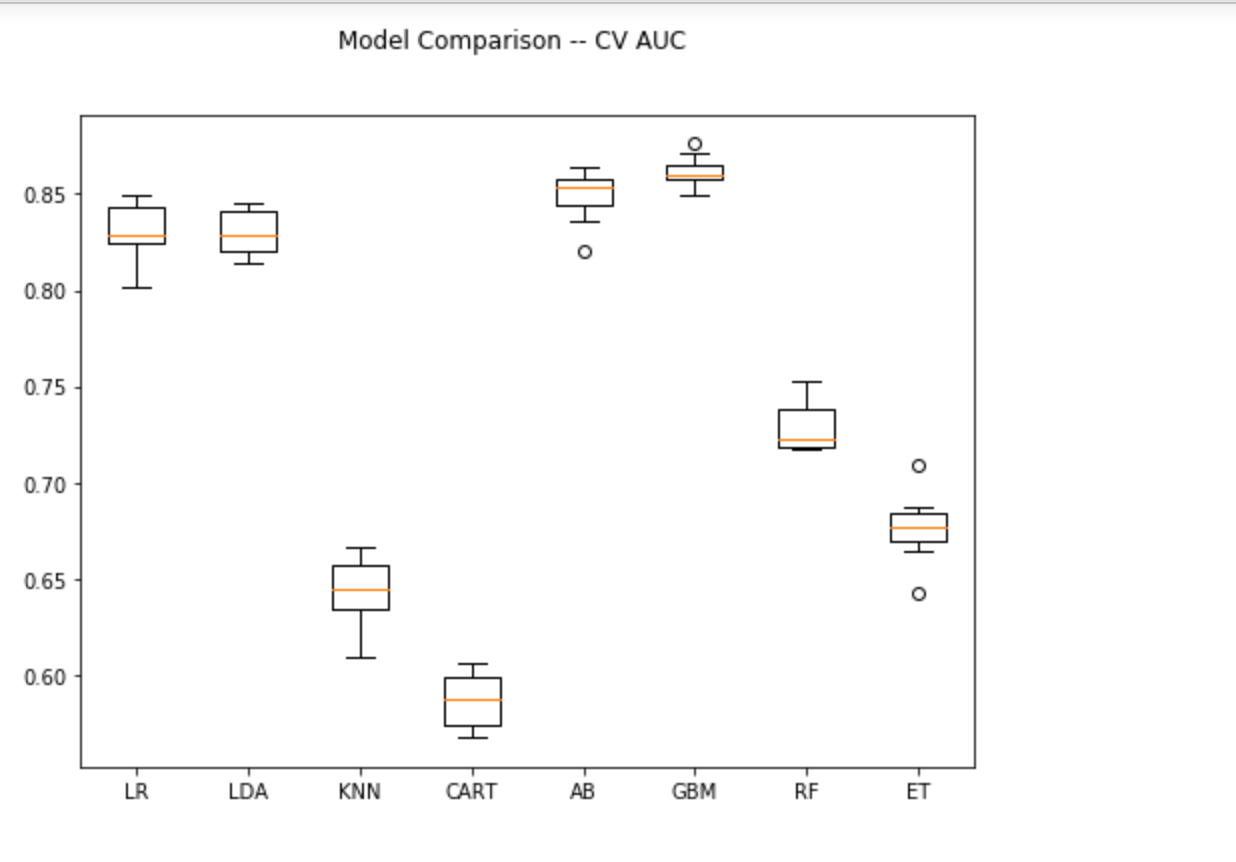
We try out different algorithms to compare and check which one is performing the best.

We select the following models:

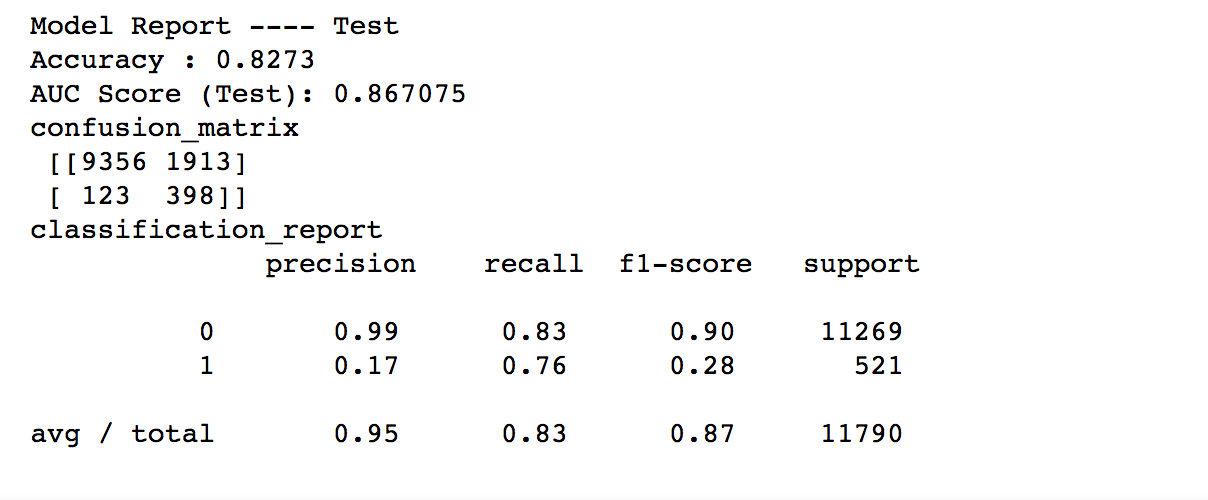
* Logistic Regression
* Linear Discriminant Analysis
* K Nearest Neighbors
* Decision Tree
* Adaptive Boosting
* Gradient Boosting
* Random Forest
* Extra Trees Classifier

We split the data into 75% training and 25% testing.





From these results, it seems like GBM is performing the best on the data.

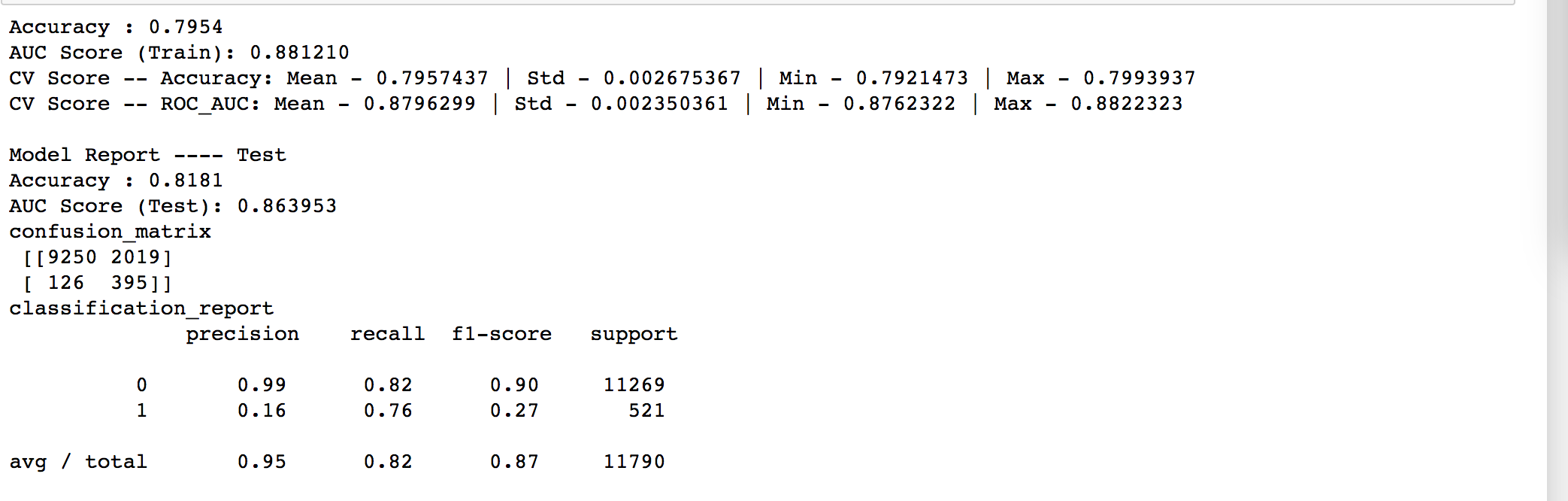


We now look at other techniques to balance the data.

# Oversampling:

We oversample the minority class to match the majority class in the dataset. This oversampling is done only on the training dataset. No such thing is done for testing dataset.

We apply GBM to the data to check if there is any improvement in the performance metrics.

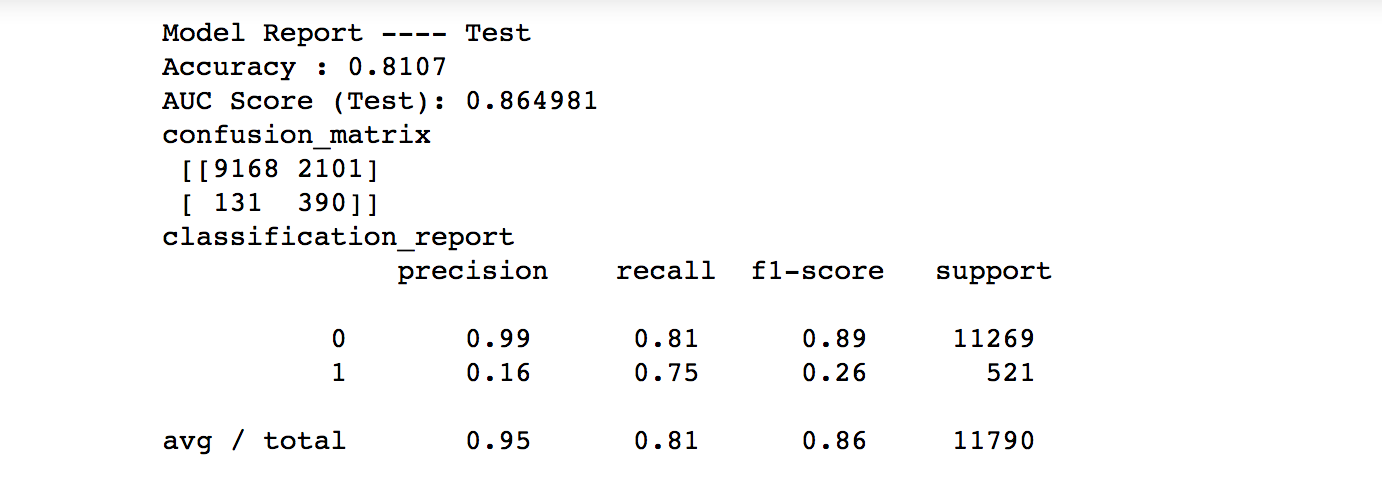


We see that there is no significant improvement using this method on the test dataset.

# Under-sampling:

We under-sample the majority class to match the minority class in the dataset. This under-sampling is done only on the training dataset. No such thing is done for testing dataset.

We apply GBM to the data to check if there is any improvement in the performance metrics.



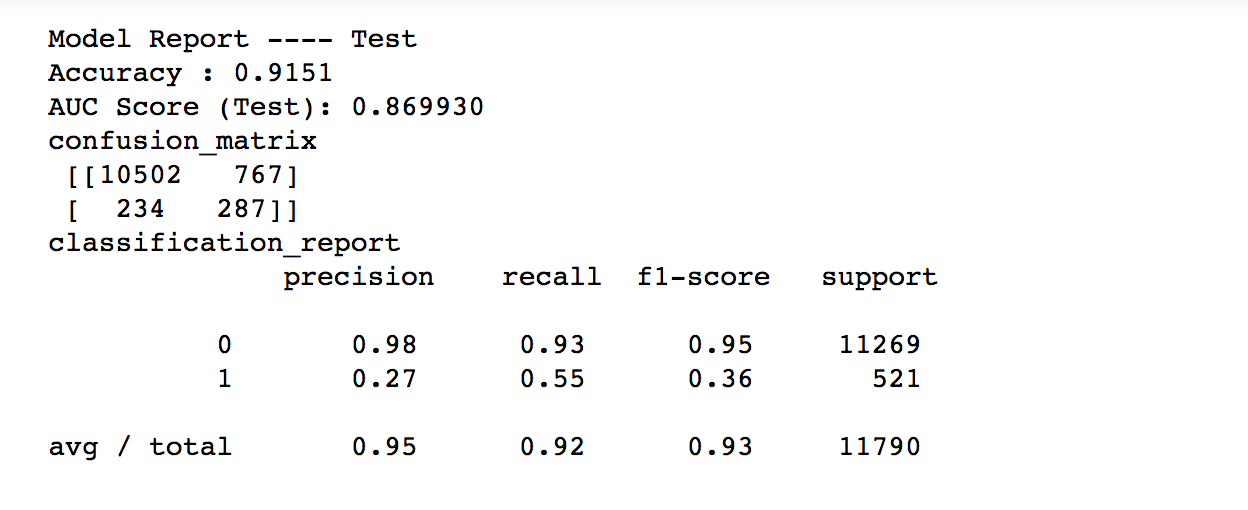
The results are similar to what we observed from the previous two methods.

# SMOTE:

The SMOTE algorithm can be broken down into following steps:

1. Randomly pick a point from the minority class.
2. Compute the k-nearest neighbors (for some pre-specified k) for this point.
3. Add k new points somewhere between the chosen point and each of its neighbors.

After balancing the data, we applied GBM to model the data.

****

While the AUC remains the same, there is a change in Precision and Recall values compared to the previous two methods.

The choice of classification depends on the cost function the user has in reference to the relative values of falsely calling out 1’s vs 0’s.

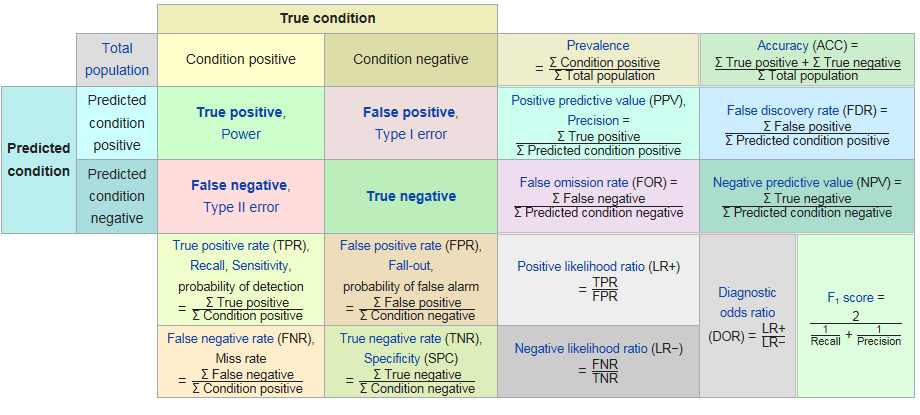
Notes on Statistical Tests:

**precision** (also called [positive predictive value](https://en.wikipedia.org/wiki/Positive_predictive_value" \o "Positive predictive value)) is the fraction of relevant instances among the retrieved instances

**recall** (also known as [sensitivity](https://en.wikipedia.org/wiki/Sensitivity_and_specificity" \o "Sensitivity and specificity)) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances

The F1 **score** is the harmonic average of the precision and recall, where an F1 **score** reaches its best value at 1 (perfect precision and recall) and worst at 0.

a **confusion matrix** is a table with two rows and two columns that reports the number of *false positives*, *false negatives*, *true positives*, and *true negatives*. This allows more detailed analysis than mere proportion of correct classifications (accuracy).



[Accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision" \l "In_binary_classification" \o "Accuracy and precision) is a weighted arithmetic mean of Precision and Inverse Precision (weighted by Bias) as well as a weighted arithmetic mean of Recall and Inverse Recall (weighted by Prevalence).[[1]](https://en.wikipedia.org/wiki/Precision_and_recall" \l "cite_note-Powers2011-1) Inverse Precision and Inverse Recall are simply the Precision and Recall of the inverse problem where positive and negative labels are exchanged (for both real classes and prediction labels). Recall and Inverse Recall, or equivalently true positive rate and false positive rate, are frequently plotted against each other as [ROC](https://en.wikipedia.org/wiki/Receiver_operating_characteristic" \o "Receiver operating characteristic) curves and provide a principled mechanism to explore operating point tradeoffs.

the application of Recall, Precision and F-measure are argued to be flawed as they ignore the true negative cell of the contingency table, and they are easily manipulated by biasing the predictions.[[1]](https://en.wikipedia.org/wiki/Precision_and_recall" \l "cite_note-Powers2011-1)

The first problem is 'solved' by using [Accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision" \l "In_binary_classification" \o "Accuracy and precision) and …..

Metrics not captured in this paper

… the second problem is 'solved' by discounting the chance component and renormalizing to [Cohen's kappa](https://en.wikipedia.org/wiki/Cohen%27s_kappa" \o "Cohen's kappa), but this no longer affords the opportunity to explore tradeoffs graphically.

However, [Informedness](https://en.wikipedia.org/wiki/Informedness" \o "Informedness) and [Markedness](https://en.wikipedia.org/wiki/Markedness" \o "Markedness) are Kappa-like renormalizations of Recall and Precision,[[3]](https://en.wikipedia.org/wiki/Precision_and_recall" \l "cite_note-3) and their geometric mean [Matthews correlation coefficient](https://en.wikipedia.org/wiki/Matthews_correlation_coefficient" \o "Matthews correlation coefficient) thus acts like a debiased F-measure.