OUD Progression – Building a Predictive Model

# Sample Selection:

Starting from MarketScan Medicaid dataset, we selected enrollees who has a clean period in 2013, meaning that the selected enrollee does not have an opioid RX during last 6 months of 2013.

Specifically, RX is a lists of code sets (icd9 and icd10) for or related to Opioid User Disorder.

In addition, the selected enrollee:

* Must have 11+ months of eligibility in each year from 2013 from 2016
* Must be 12+ years old as start of Measurement Year 2014 (1/1/2014)
* All members in Pharmacy, Medical Claims and Admissions will have an eligibility record for each year

# 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 |
| 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 | PCP (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 PCP 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 | Fentanyl 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 | Physician who ordered Rx |
| pharmid | Pharmacy that filled Rx |
| daysupp | # days supply of Rx |
| 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: 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 | At least 1 antidepressant Rx during year |
| 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:

Our aim is to predict whether a patient will be Opioid Dependent or not based on the data of that patient for the previous 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.

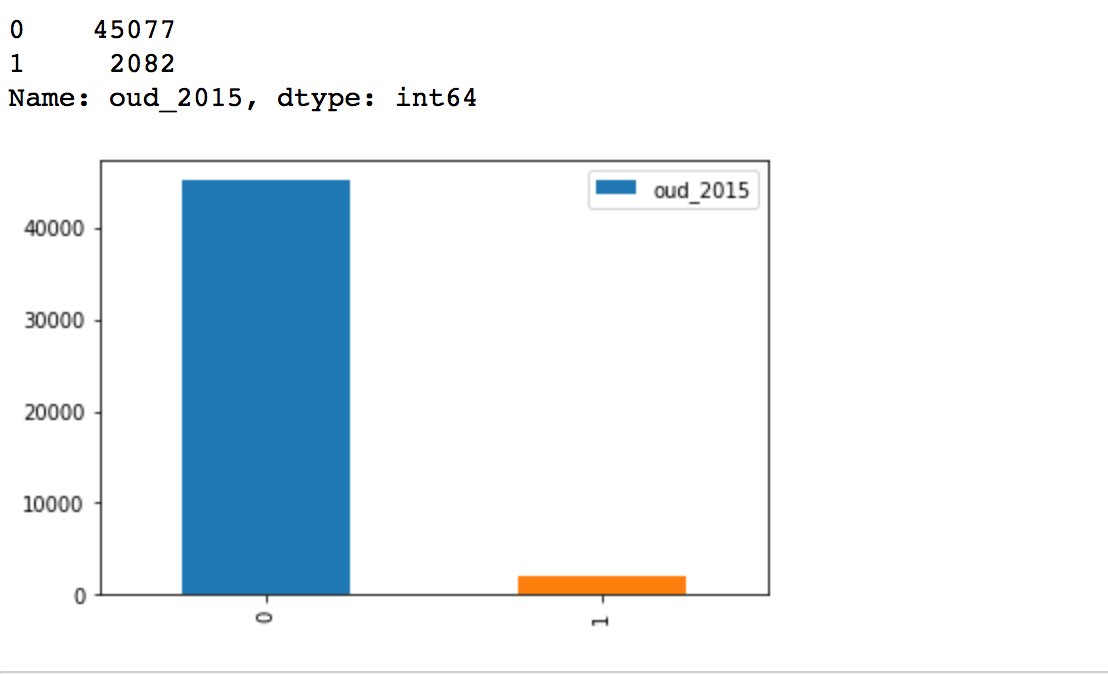
We then add the OUD diagnosis from med\_15 file to the dataset for the common ENROLIDs.

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.

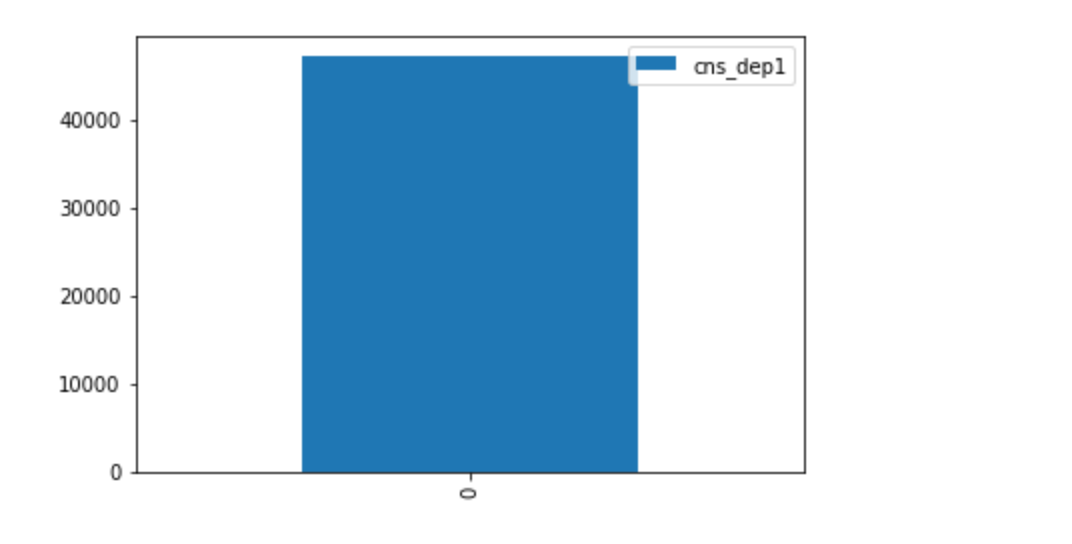
# Exploratory Data Analysis:

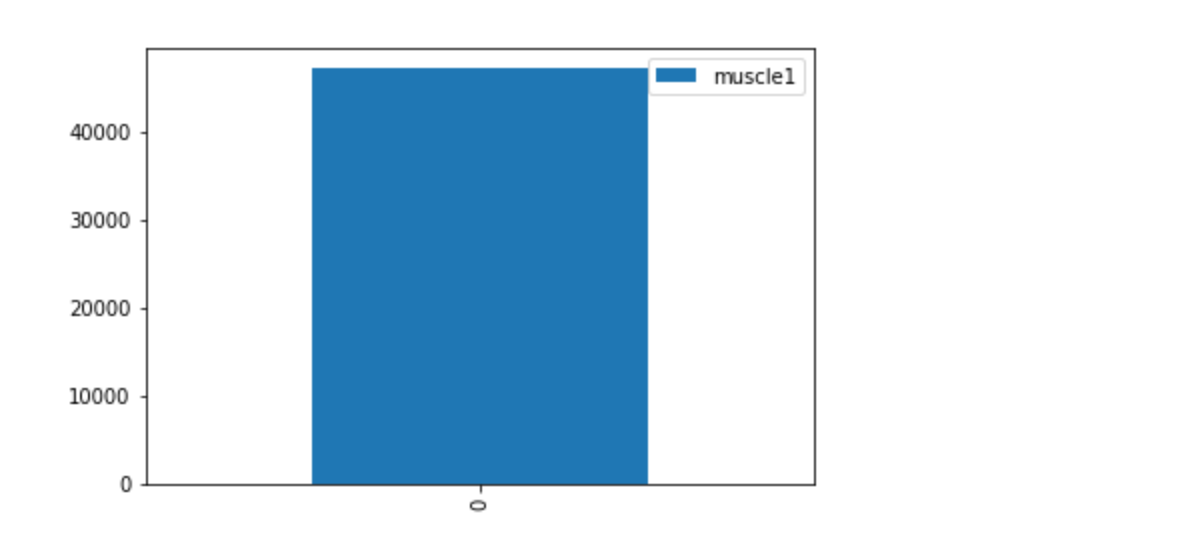
We analyze the data using graphs.

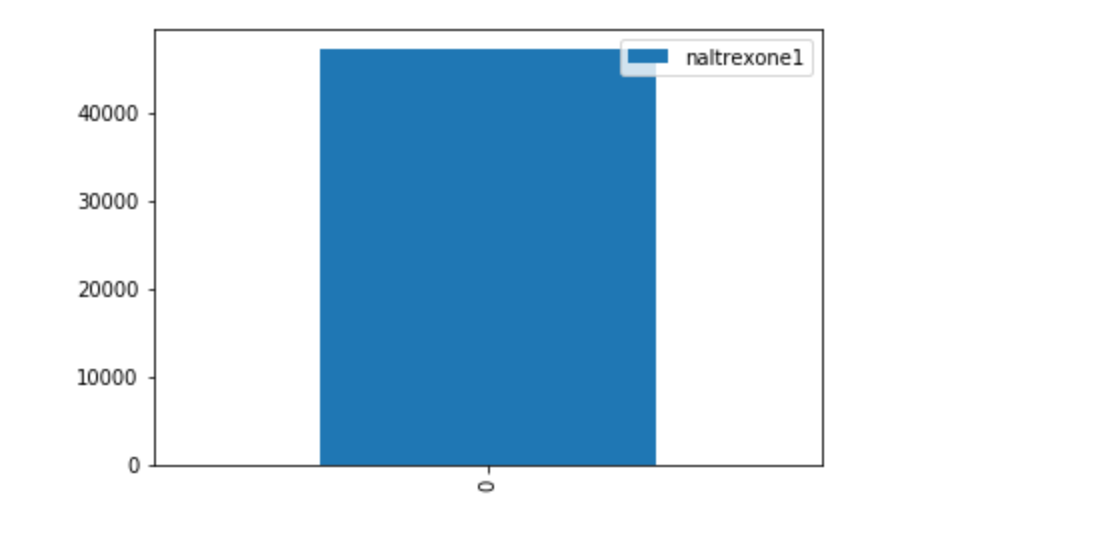
The first thing we observe is that the data is highly imbalanced.



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.



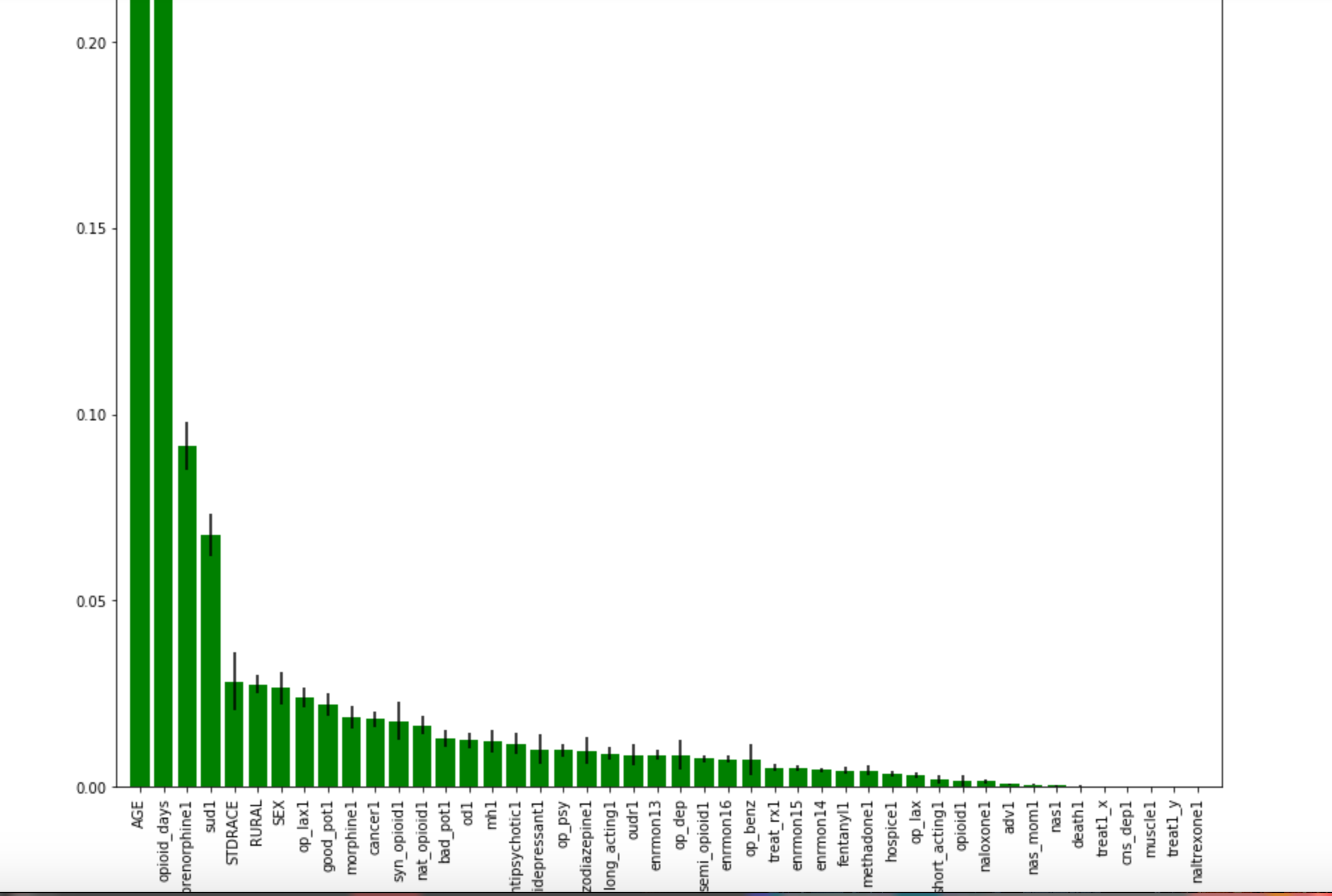


****

**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.

****

# 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. Oversample the minority class to match the majority class.
2. Undersample 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.

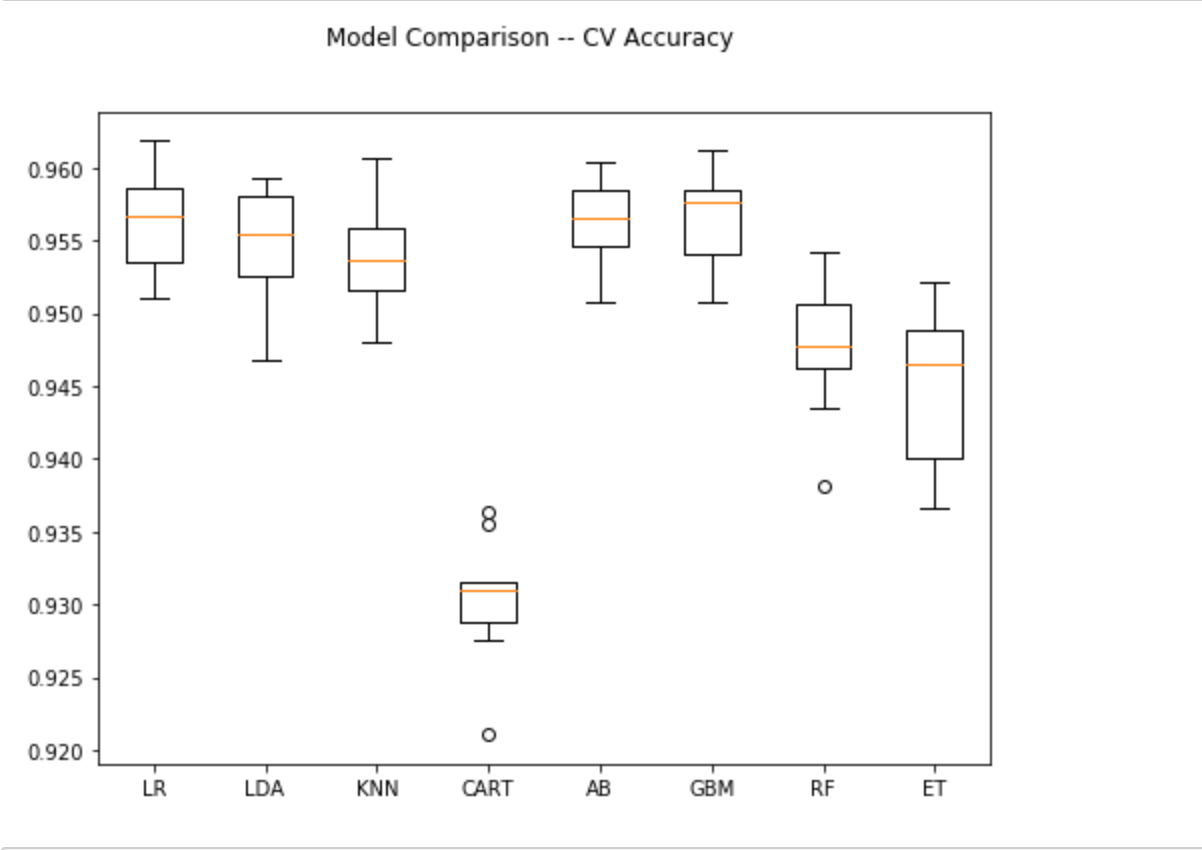
# Using Class Weights:

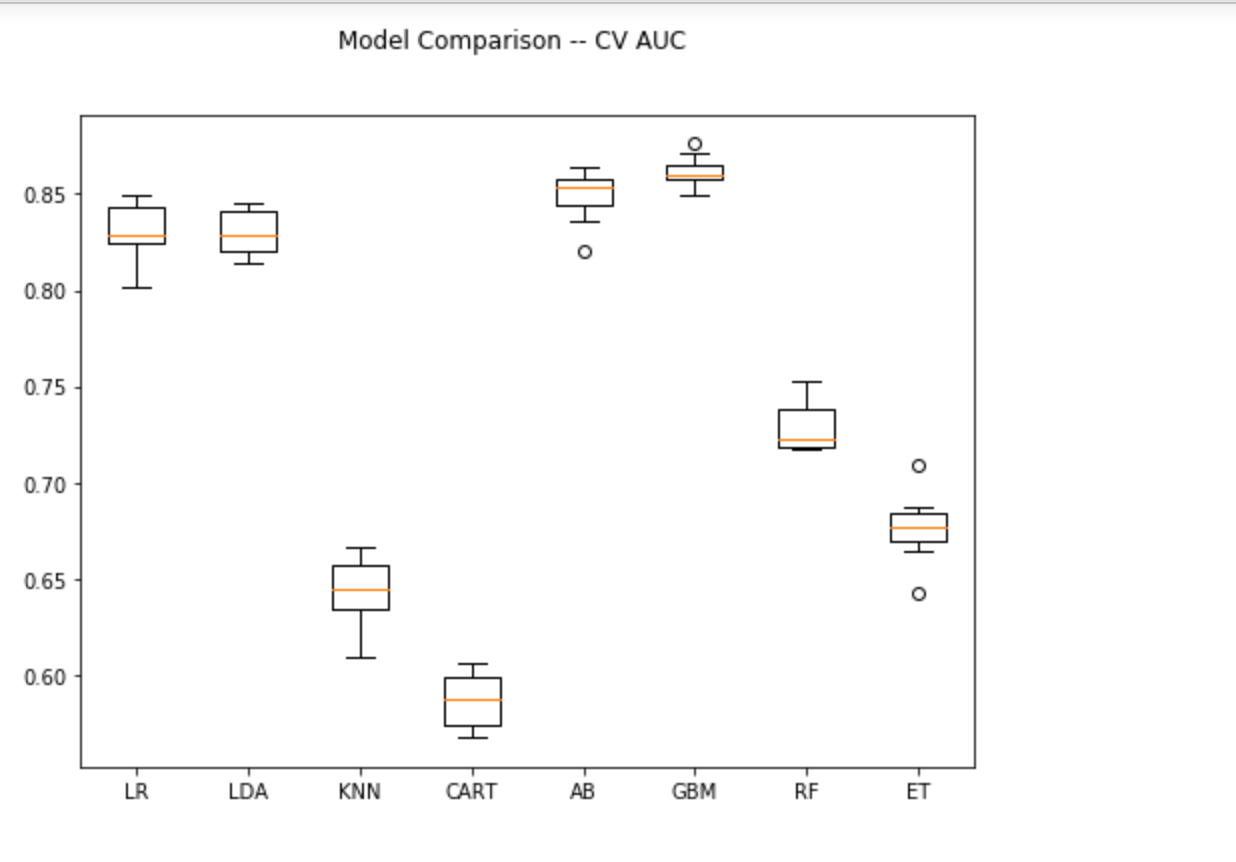
We try out different algorithms to compare and check which one is performing the best.

We select the following models:

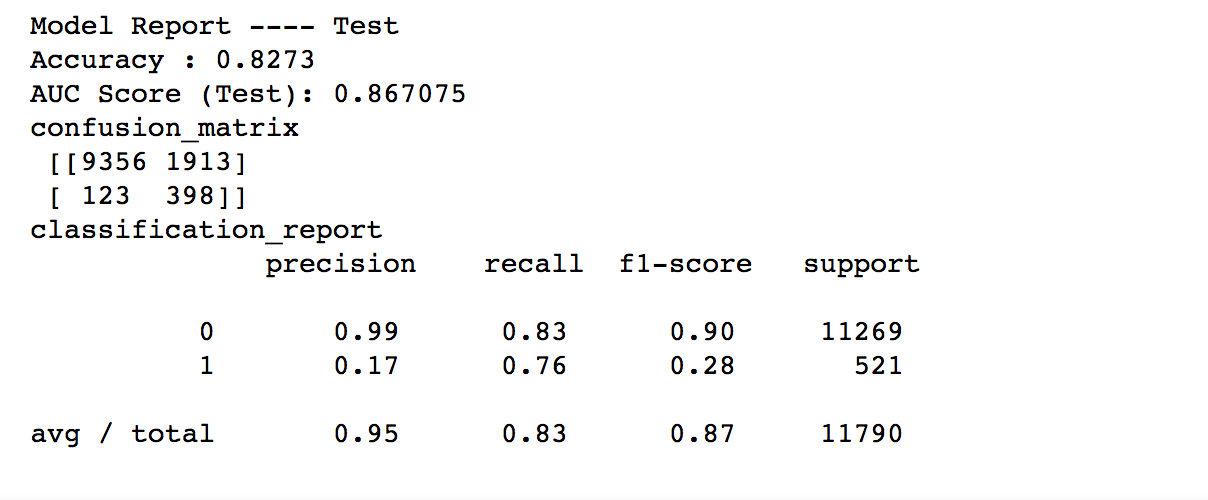
* Logistic Regression
* Linear Discriminant Analysis
* K Nearest Neighbors
* Decision Tree
* Adaboost
* Gradient Boosting
* Random Forest
* Extra Trees Classifier

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





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

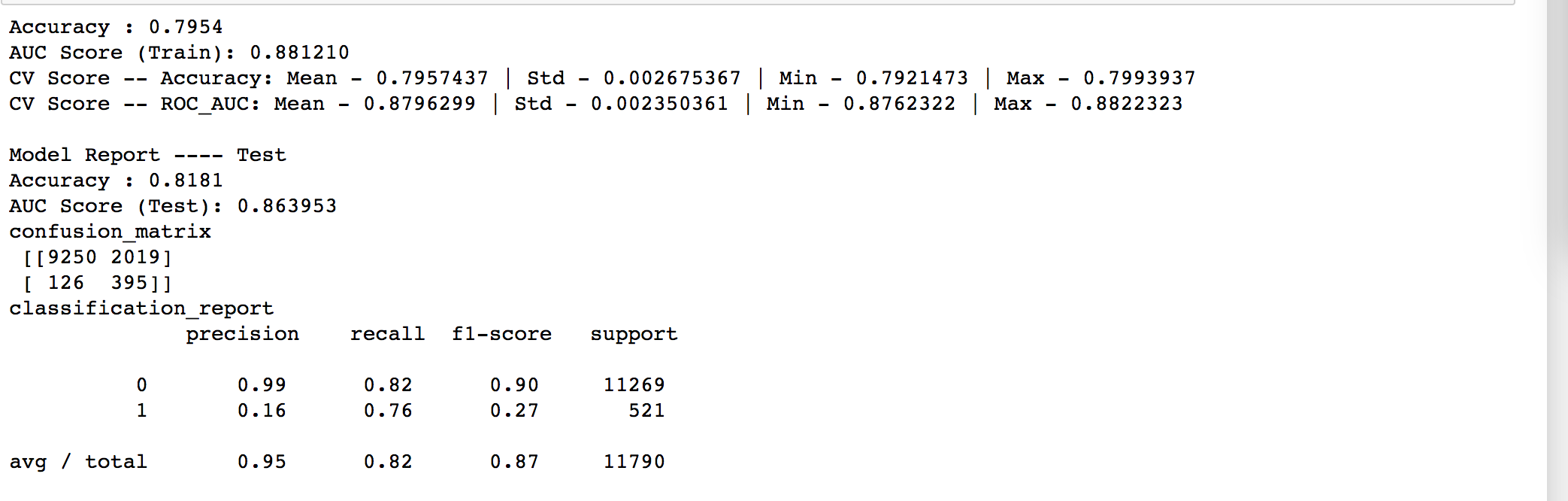


We now look at other techniques to balance the data.

# 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.

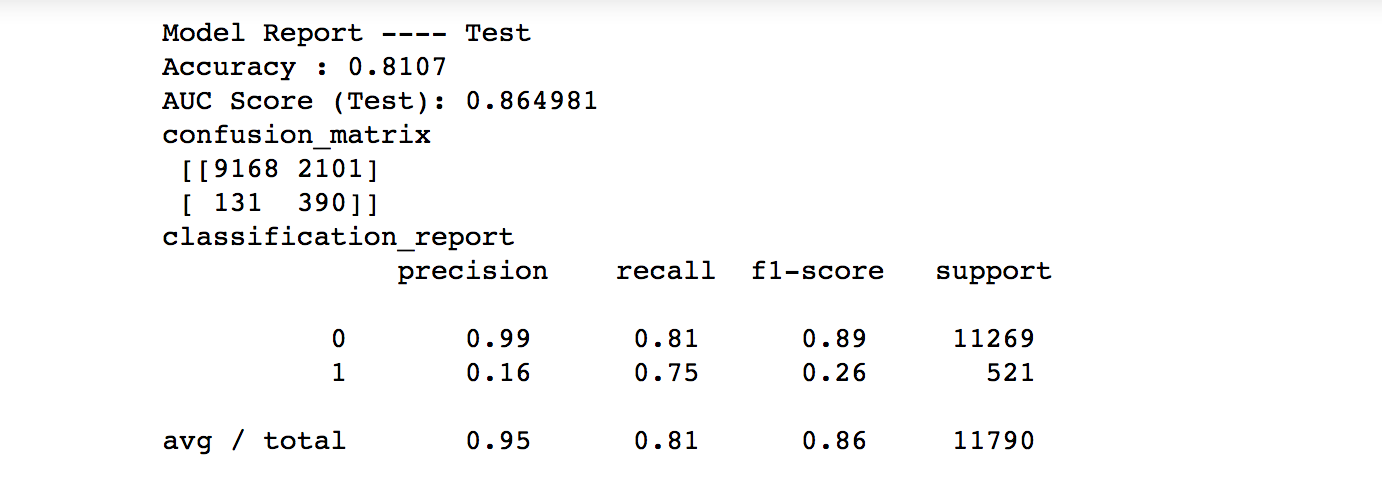


We see that there is no significant improvement using this method on the test dataset.

# Undersampling:

We undersample the majority class to match the minority 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.



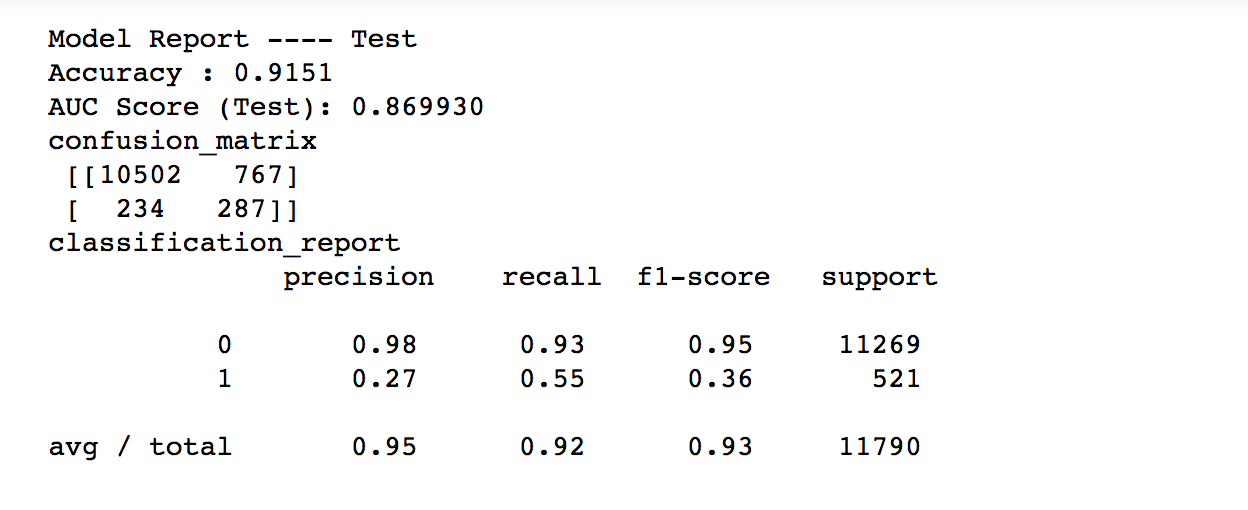
The results are similar to what we observed from the previous two methods.

# 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.

It depends on the cost function to choose the method of classification.