OPIATE PRESCRIPTIONS & OVERDOSES

PROJECT TEAM:	ROY SUBHADEEP (A20358508),
	BOTHARA SHUBHAM (A20383550)
INSTRUCTOR:	ADAM M. MCELHINNEY
COURSE:	CSP-571 DATA PREPARATION AND ANALYSIS
SCHOOL:	ILLINOIS INSTITUTE OF TECHNOLOGY, CHICAGO

MOTIVATION AND BACKGROUND

Over the past 15 years in United States, deaths from opioid prescriptions have increased four folds.

Daily approximately 113 people die in U.S. due to drug overdoses and nearly half of the yearly deaths occur due to overdoses. There were more than 38000 deaths in US in 2010 and which also exceeded the deaths by motor accidents in 2011. This has remained a serious concern over the years and CDC (Centre for Disease Control & Prevention) characterized this as an epidemic in 2012 with most deaths deemed preventable.

The massive increase in largely due to the use of opioids as pain relievers while the level of pain experienced by most Americans remains largely unchanged.

Intuitively it follows that unnecessary prescriptions play a significant role in drug overdoses. Patients who do not need it, run the risk of becoming addicted themselves or they are tempted to sell their refills for significant profits, increasing the supply of narcotics in the illegal drug market.

Over the years this has been a very active area of research. An effective strategy in identifying instances of overdoses can be a life-saving endeavour.

In the end, the goal remains to use machine learning techniques to predict likelihood that a given prescriber is a prescriber of opiates.

SCOPE OF ANALYSIS

This is a very active area of research and there are plenty of studies to perform. However, the scope of this document remains restricted primarily to detecting sources of significant quantities of opioids and other statistical findings. This study is restricted to the below analysis.

- Detecting primary sources of opiate prescriptions.
- Statistical findings on demography for patterns and trends.
- Detection of important drugs, prescriber's specialty and other important features contributing more toward the prediction.
- Examine overdose deaths in race and age groups.

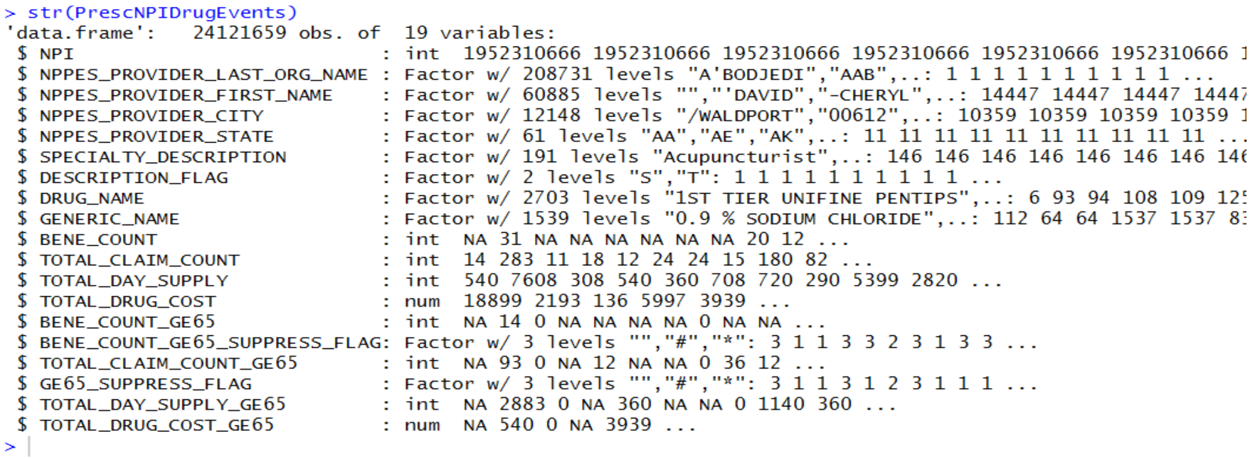
DATA EXTRACTION

All the data is available under public domain license and can be downloaded from CMS (Centre of Medicare and Medicaid services) and CDC (Centre for Disease Control and Prevention) websites.

- Prescriber Data: The prescriber data is downloaded from https://www.cms.gov/. The available data is only for the year 2013 and 2014. For this study only 2014 data has been used. The 2014 zip file contains data spread across two files, drug information and prescriber metadata.

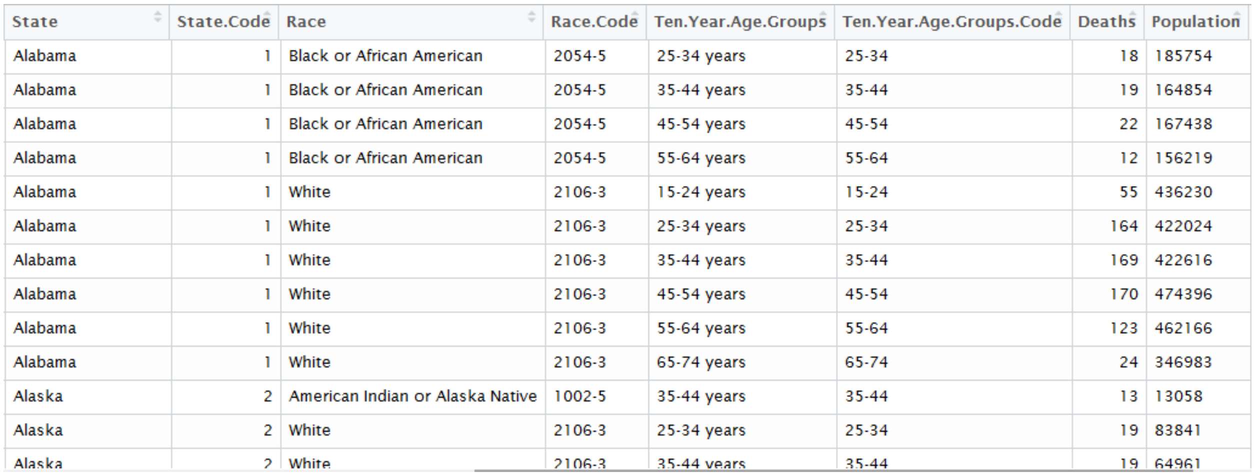
  The file PartD_Prescriber_PUF_NPI_14.txt contains prescriber metadata and PartD_Prescriber_PUF_NPI_Drug_14.txt contains the drugs prescribed by the prescribers under Medicare Part D plan.

  The total size of the dataset is 2.5 GB. The data which is available is in long format. Each prescriber Id (NPI) is spread across multiple rows based on the drugs that they have prescriber. Below is the structure of the file.



```
> str(PrescNPIDrugEvents)
'data.frame':	24121659 obs. of  19 variables:
 $ NPI                           : int  1952310666 1952310666 1952310666 1952310666 1952310666 1952310666 1
 $ NPPES_PROVIDER_LAST_ORG_NAME  : Factor w/ 208731 levels "A'BODJEDI","AAB",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ NPPES_PROVIDER_FIRST_NAME     : Factor w/ 60885 levels "","'DAVID","-CHERYL",..: 14447 14447 14447 1444
 $ NPPES_PROVIDER_CITY           : Factor w/ 12148 levels "/WALDPORT","00612",..: 10359 10359 10359 10359 1
 $ NPPES_PROVIDER_STATE          : Factor w/ 61 levels "AA","AE","AK",..: 11 11 11 11 11 11 11 11 11 11 ...
 $ SPECIALTY_DESCRIPTION         : Factor w/ 191 levels "Acupuncturist",..: 146 146 146 146 146 146 146 14
 $ DESCRIPTION_FLAG              : Factor w/ 2 levels "S","T": 1 1 1 1 1 1 1 1 1 1 ...
 $ DRUG_NAME                     : Factor w/ 2703 levels "1ST TIER UNIFINE PENTIPS",..: 6 93 94 108 109 12
 $ GENERIC_NAME                  : Factor w/ 1539 levels "0.9 % SODIUM CHLORIDE",..: 112 64 64 1537 1537 8
 $ BENE_COUNT                    : int  NA 31 NA NA NA NA NA NA 20 12 ...
 $ TOTAL_CLAIM_COUNT             : int  14 283 11 18 12 24 24 15 180 82 ...
 $ TOTAL_DAY_SUPPLY              : int  540 7608 308 540 360 708 720 290 5399 2820 ...
 $ TOTAL_DRUG_COST               : num  18899 2193 136 5997 3939 ...
 $ BENE_COUNT_GE65               : int  NA 14 0 NA NA NA NA 0 NA NA ...
 $ BENE_COUNT_GE65_SUPPRESS_FLAG: Factor w/ 3 levels "","#","*": 3 1 1 3 3 2 3 1 3 3 ...
 $ TOTAL_CLAIM_COUNT_GE65        : int  NA 93 0 NA 12 NA NA 0 36 12 ...
 $ GE65_SUPPRESS_FLAG            : Factor w/ 3 levels "","#","*": 3 1 1 3 1 2 3 1 1 1 ...
 $ TOTAL_DAY_SUPPLY_GE65         : int  NA 2883 0 NA 360 NA NA 0 1140 360 ...
 $ TOTAL_DRUG_COST_GE65          : num  NA 540 0 NA 3939 ...
> 
```

- Drug Category information: This dataset is also available in the CMS website. This data classifies drug names into four categories, opioids, antipsychotics, antibiotics and high risk medications.
- Demography Data: Demography data can be downloaded from CDC website https://wonder.cdc.gov/controller/datarequest/D76. This downloaded data has the number of deaths per population for drug/alcohol induced cases in 2014 and has been grouped under state, race and age group. However, there are many other ways it can be downloaded.



State	State.Code	Race	Race.Code	Ten.Year.Age.Groups	Ten.Year.Age.Groups.Code	Deaths	Population
Alabama	1	Black or African American	2054-5	25-34 years	25-34	18	185754
Alabama	1	Black or African American	2054-5	35-44 years	35-44	19	164854
Alabama	1	Black or African American	2054-5	45-54 years	45-54	22	167438
Alabama	1	Black or African American	2054-5	55-64 years	55-64	12	156219
Alabama	1	White	2106-3	15-24 years	15-24	55	436230
Alabama	1	White	2106-3	25-34 years	25-34	164	422024
Alabama	1	White	2106-3	35-44 years	35-44	169	422616
Alabama	1	White	2106-3	45-54 years	45-54	170	474396
Alabama	1	White	2106-3	55-64 years	55-64	123	462166
Alabama	1	White	2106-3	65-74 years	65-74	24	346983
Alaska	2	American Indian or Alaska Native	1002-5	35-44 years	35-44	13	13058
Alaska	2	White	2106-3	25-34 years	25-34	19	83841
Alaska	2	White	2106-3	35-44 years	35-44	19	64961

DATA FORMATTING & CLEANING

FORMATTING

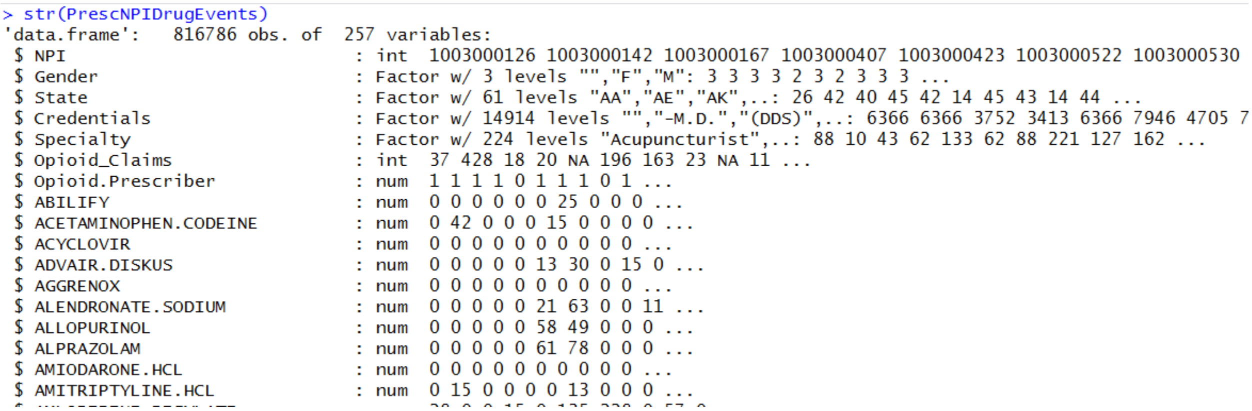
The downloaded data is merged and formatted before analysis and modelling. The prescriber information in long format is converted into wide format so that NPI is the primary key and all the related drug prescriptions are now split into multiple columns. Then the prescriber metadata is merged to the main dataset to add some details about prescriber information.

The top 250 drugs out of 2700 drugs are selected based on the quantity prescribed. This subset is majorly done to reduce the computational overhead due to hardware limitations. The scope was intended to expand but it did not go very well with the individual pc's.

A binary target label is created if the prescriber has prescribed even one opioid drug or not. The final format of the main dataset is 257 columns where 250 of them are drugs prescribed, 6 columns are prescriber information and one target label.

The demography data is filtered on drug induced cases and alcohol related death rows are removed. Only columns of importance and kept. Basic cleansing operations like removing extra statistics and NA values are performed.

R Libraries Used: deplyr, tidyr

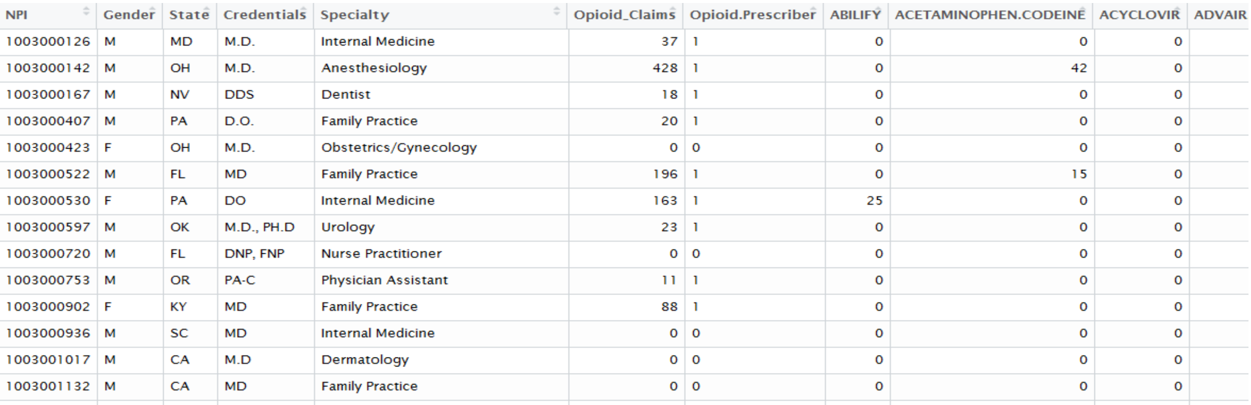


```
> str(PrescNPIDrugEvents)
'data.frame':	816786 obs. of  257 variables:
 $ NPI                               : int  1003000126 1003000142 1003000167 1003000407 1003000423 1003000522 1003000530
 $ Gender                            : Factor w/ 3 levels "","F","M": 3 3 3 3 2 3 2 3 3 3 ...
 $ State                             : Factor w/ 61 levels "AA","AE","AK",..: 26 42 40 45 42 14 45 43 14 44 ...
 $ Credentials                       : Factor w/ 14914 levels "","-M.D.","(DDS)",..: 6366 6366 3752 3413 6366 7946 4705 7
 $ Specialty                         : Factor w/ 224 levels "Acupuncturist",..: 88 10 43 62 133 62 88 221 127 162 ...
 $ Opioid_Claims                     : int  37 428 18 20 NA 196 163 23 NA 11 ...
 $ Opioid.Prescriber                 : num  1 1 1 1 0 1 1 1 0 1 ...
 $ ABILIFY                           : num  0 0 0 0 0 0 25 0 0 0 ...
 $ ACETAMINOPHEN.CODEINE             : num  0 42 0 0 0 15 0 0 0 0 ...
 $ ACYCLOVIR                         : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ADVAIR.DISKUS                     : num  0 0 0 0 0 13 30 0 15 0 ...
 $ AGGRENOX                          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ALENDRONATE.SODIUM                : num  0 0 0 0 0 21 63 0 0 11 ...
 $ ALLOPURINOL                       : num  0 0 0 0 0 58 49 0 0 0 ...
 $ ALPRAZOLAM                        : num  0 0 0 0 0 61 78 0 0 0 ...
 $ AMIODARONE.HCL                    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ AMITRIPTYLINE.HCL                 : num  0 15 0 0 0 0 13 0 0 0 ...
```

DATA CLEANING

The following summarizes the cleaning operations performed on the dataset.

- Fixing 61 states and incorrect abbreviations.
  - Correct state codes and abbreviations are web scraped using rvest from https://www.infoplease.com/state-abbreviations-and-state-postal-codes.
  - State names are manually fixed in the demography data and matched to produce correct abbreviations.
  - State codes are matched and any incorrect state codes are changed to “other” category.
- Opioid drug columns are removed from the main dataset. This is done as it makes the prediction very obvious. Not many columns are removed indicating the top drugs are not opioids.
- Columns are converted into appropriate types.
- 24% of the opioid claim columns has NA values. Since claims are the drugs prescribed including refills, it is safe to assume the NA values to be 0.
- One gender is missing. The same row had a miss in specialty as well and hence it is dropped.
- 3.12% credentials are missing. We set it as NA as we already have an others category.
- Credential columns contain a lot of junk and inconsistent values. Fixing this is not possible as there are over 500 factors probably due to errors. However, this is column can be compensated by Specialty column which is much cleaner and carry the same importance as credentials.
- Any drug columns totalling to 0 amount is rechecked for removal.

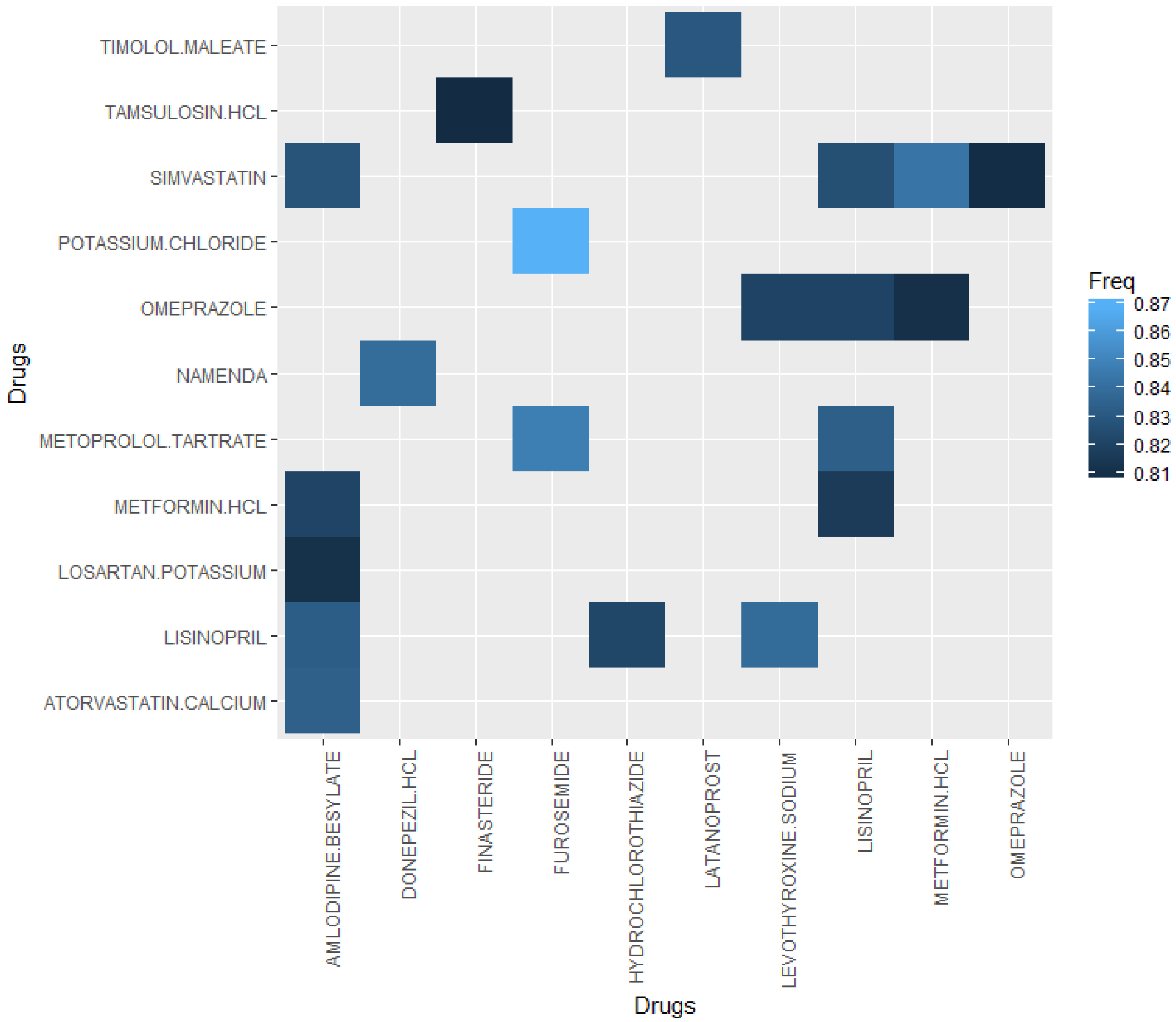


NPI	Gender	State	Credentials	Specialty	Opioid_Claims	Opioid.Prescriber	ABILIFY	ACETAMINOPHEN.CODEINE	ACYCLOVIR	ADVAIR
1003000126	M	MD	M.D.	Internal Medicine	37	1	0	0	0	
1003000142	M	OH	M.D.	Anesthesiology	428	1	0	42	0	
1003000167	M	NV	DDS	Dentist	18	1	0	0	0	
1003000407	M	PA	D.O.	Family Practice	20	1	0	0	0	
1003000423	F	OH	M.D.	Obstetrics/Gynecology	0	0	0	0	0	
1003000522	M	FL	MD	Family Practice	196	1	0	15	0	
1003000530	F	PA	DO	Internal Medicine	163	1	25	0	0	
1003000597	M	OK	M.D., PH.D	Urology	23	1	0	0	0	
1003000720	M	FL	DNP, FNP	Nurse Practitioner	0	0	0	0	0	
1003000753	M	OR	PA-C	Physician Assistant	11	1	0	0	0	
1003000902	F	KY	MD	Family Practice	88	1	0	0	0	
1003000936	M	SC	MD	Internal Medicine	0	0	0	0	0	
1003001017	M	CA	M.D	Dermatology	0	0	0	0	0	
1003001132	M	CA	MD	Family Practice	0	0	0	0	0	

EXPLORATORY DATA ANALYSIS

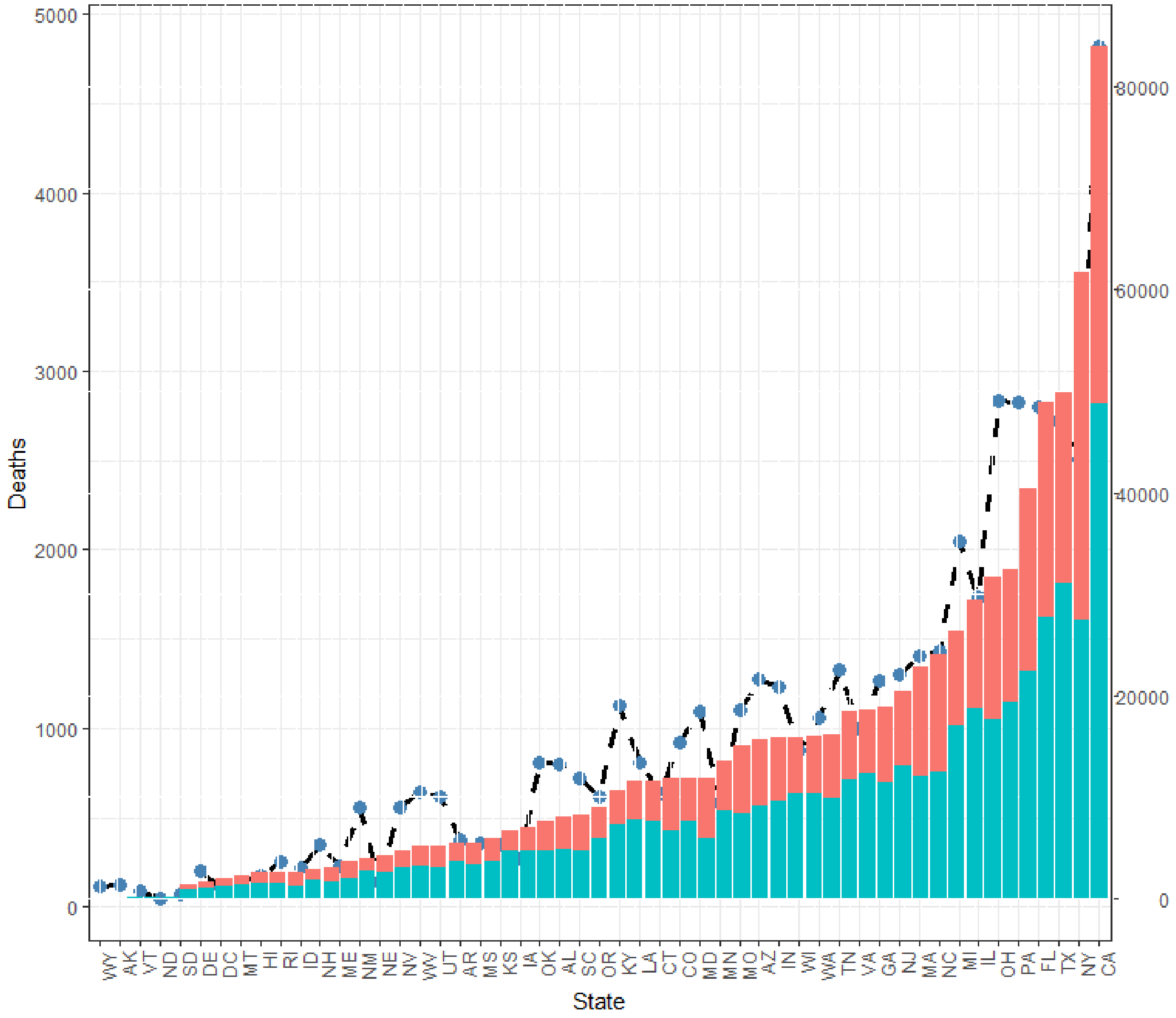
Exploratory analysis of the data was performed to understand the relationships within data.

CORRELATION PLOT



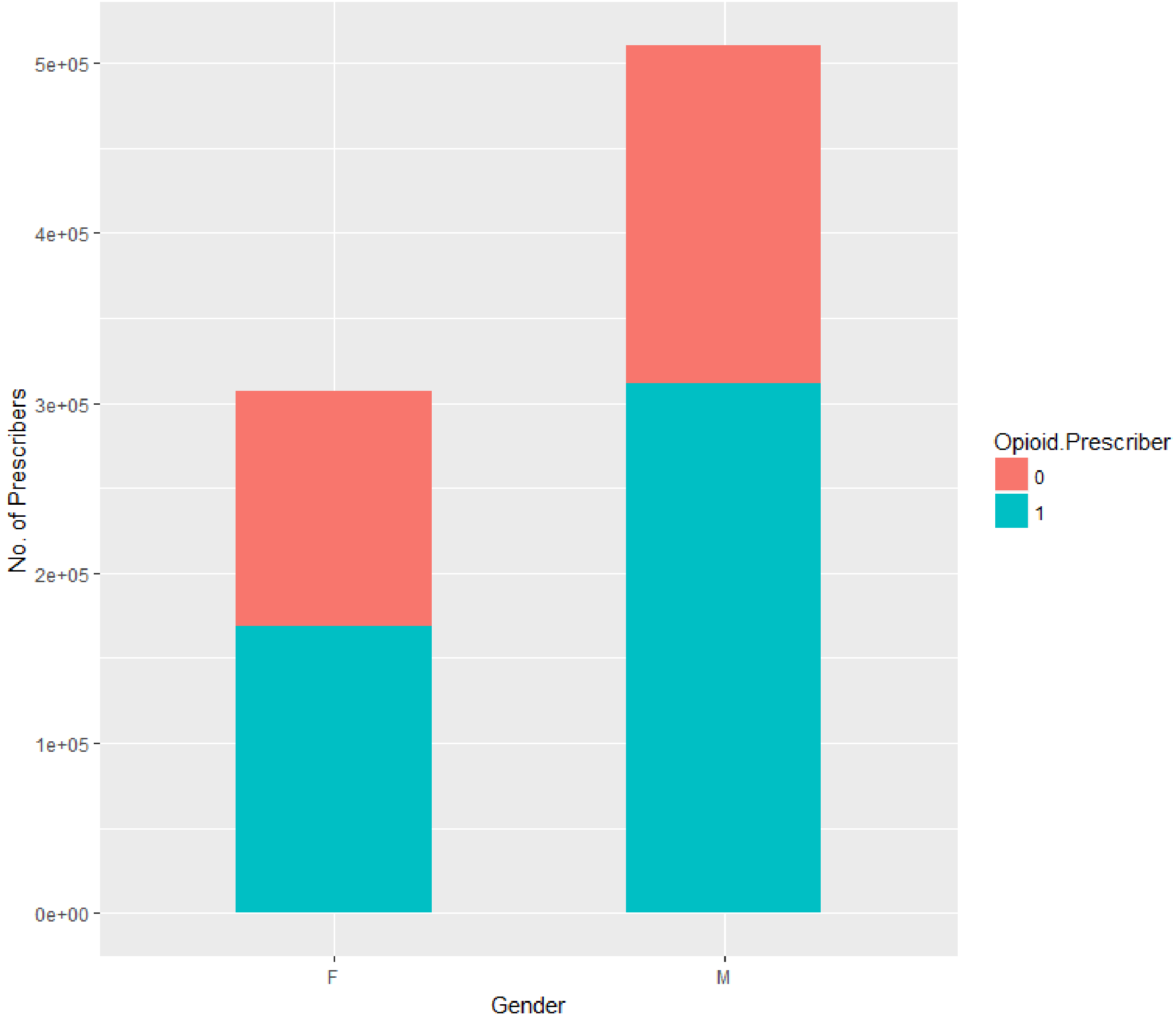
There is a high correlation between Potassium.Chloride and Furosemide. This combination is usually prescribed to treat fluid retention (Edema) in people with congestive heart failure, liver disease. Similarly, other drugs too are prescribed in critical situations only, as they have big side-effects.

BAR PLOT



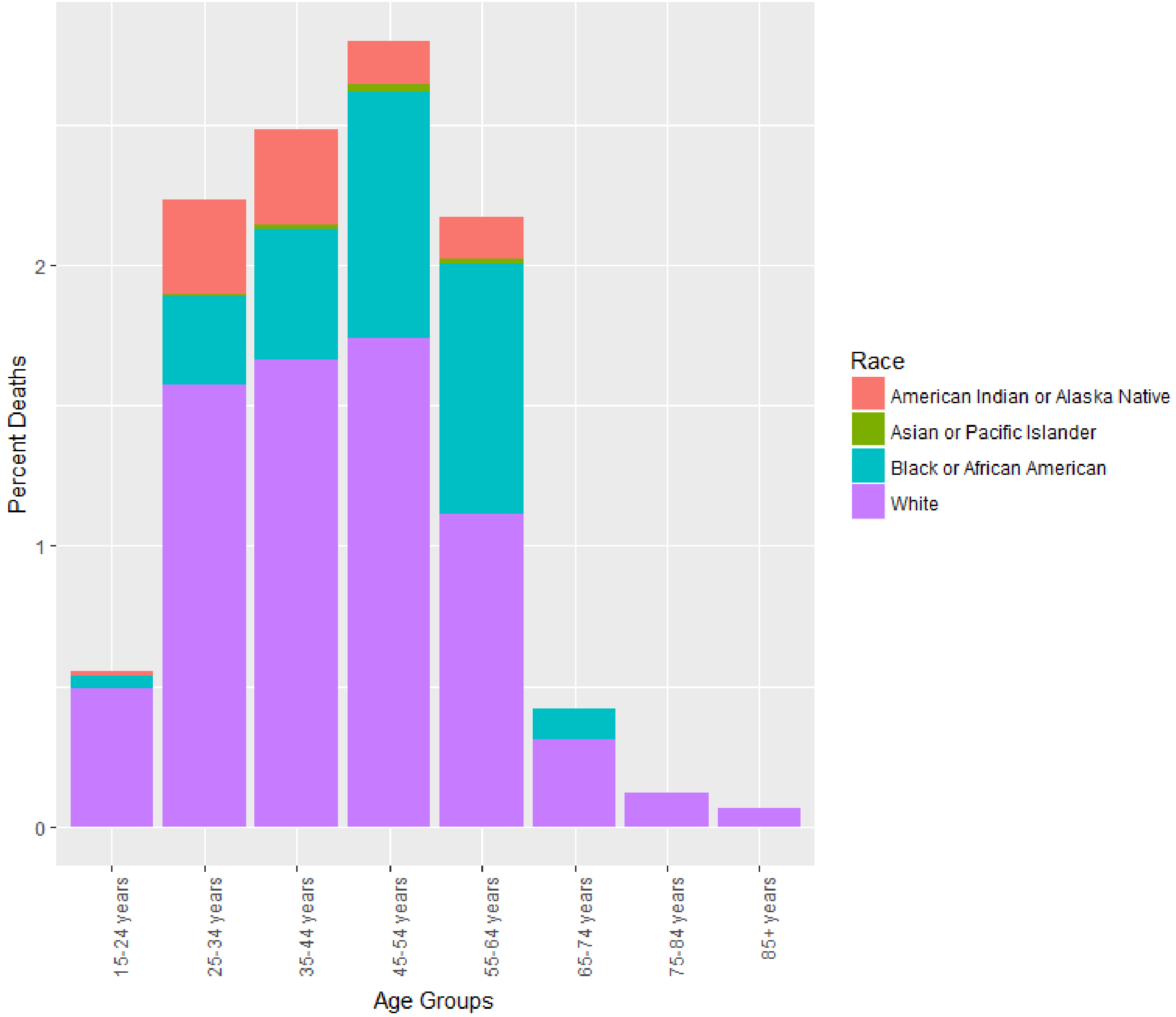
The bar chart shows the number of prescribers in a state, filled with opioid(blue) and non-opioid(red) prescribers with its y-axis on the right side. The line plot shows the number of opioid related deaths according to states with its y-axis on the left side. It can be seen from the plot that there exists a strong relationship between opioid related deaths and opioid prescription in states, with higher correlation in California, Texas, New Mexico Florida.

BAR PLOT



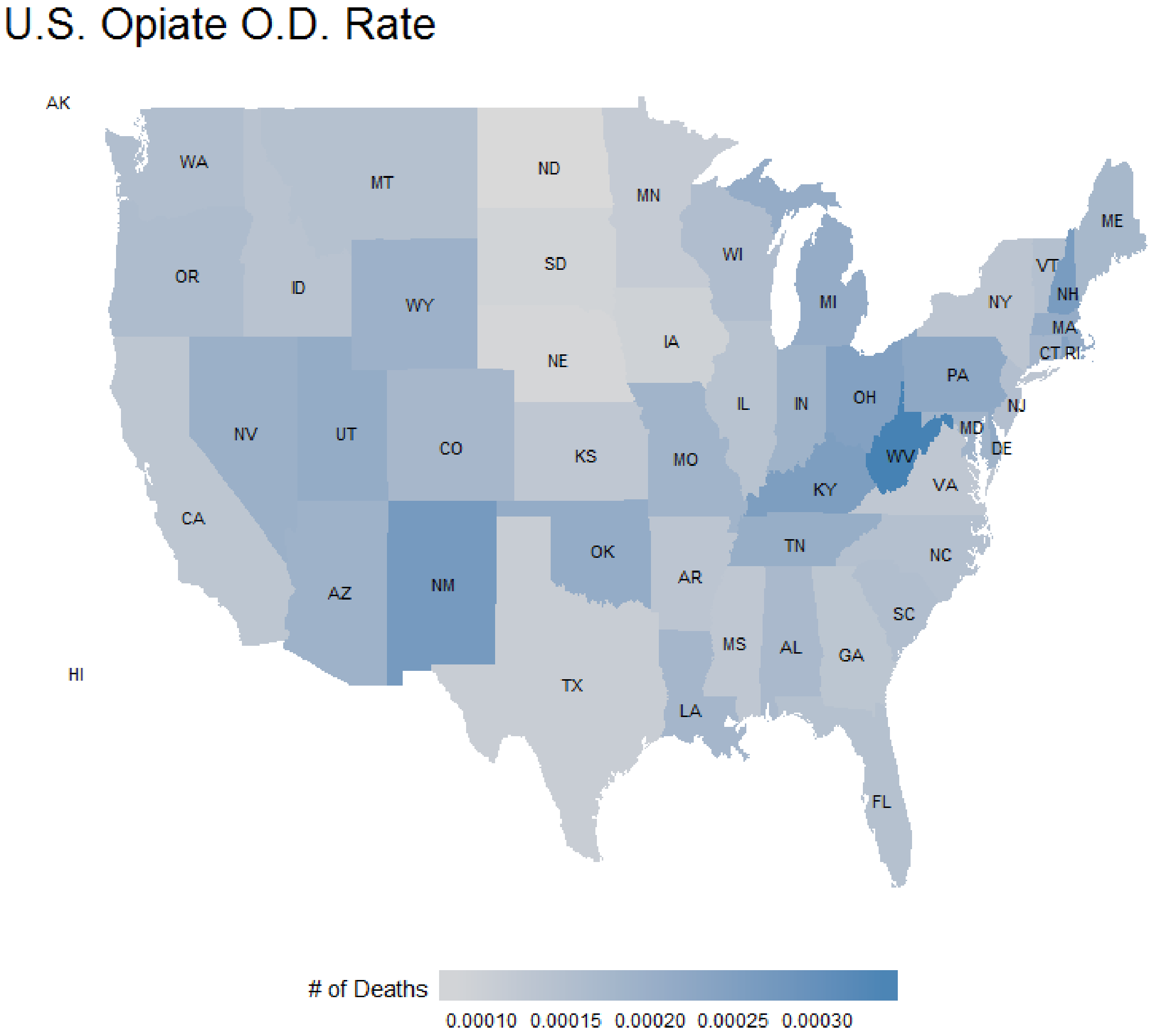
The bar chart plots the relationship between the gender of prescriber and their prescriptions. The ratio of the number of Opioid Prescribers to non-opioid prescribers seems to be higher in males. Looks like more number of males are inclined towards prescribing opioids.

BAR PLOT



To understand the relationship between age, race and opioid related deaths, we plotted the above graph. From the above bar chart, we can say that across all age groups there seems to be higher opioid related deaths in white people. The age group 15-24 years have higher percentage of white deaths while non-white deaths cease to exist beyond 75 years.

RMAP PLOT



Plotting the death per capita on the state to analyse if there exists a relationship between the geographical location of state and opioid related deaths. There seems to be higher Opioid related deaths percent in the boundary and closer to boundary states with exception to North Dakota. Also, there seems to be stronger correlation in the state of New Mexico and West Virginia.

WORDCLOUD

PREDICTIVE MODELLING TECHNIQUES

DATA TRANSFORMATION

ONE-HOT ENCODING:

Few categorical columns like specialty and states can be of paramount importance when it comes to predicting opioid prescriptions and therefore must be scored separately. 50 states and around 250 specialties are binary encoded using dummies library in R. Creating a dummy was preferred over creating a model matrix as it retains the convenience of a data frame. The final data set after transforming now contains 496 columns.

STRATEFIED SAMPLING:

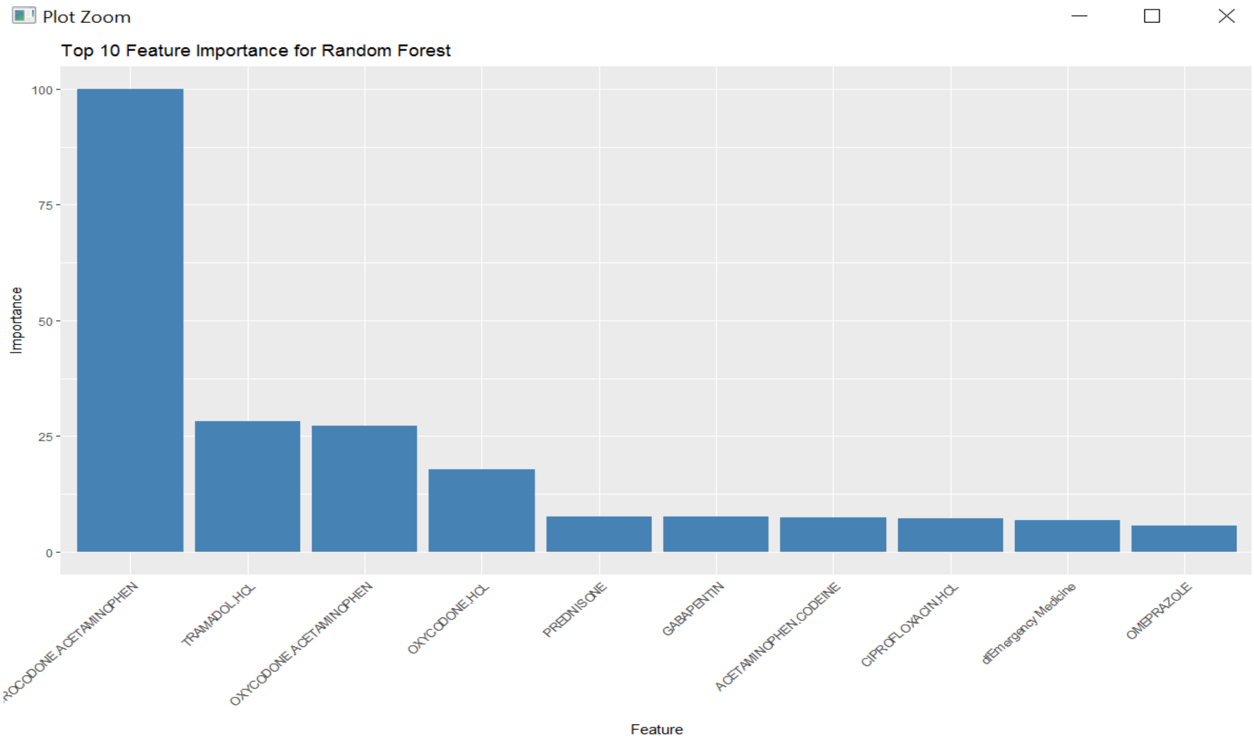
Data is partitioned into 80-20 train test datasets using caret stratified sampling.

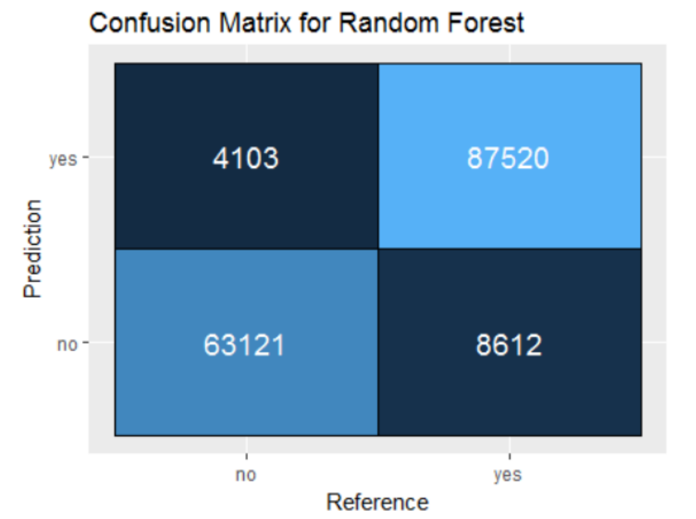
MACHINE LEARNING MODELS

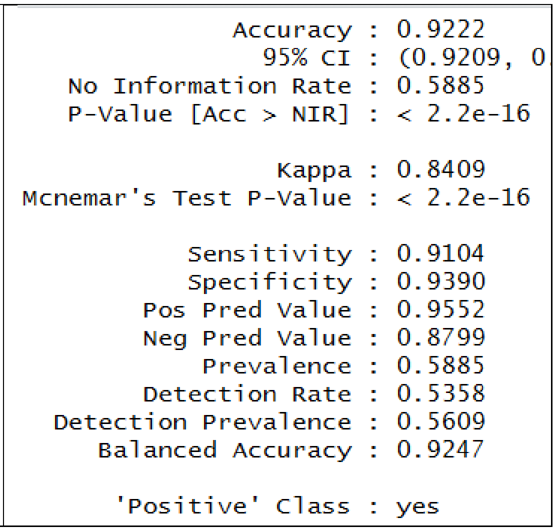
Six machine learning models are tried out and ensembled later for better performance.

RANDOM FOREST:

Random forest is optimized using caret train control cross validation using 5 folds/no-repeat and grid tuned. This is computationally expensive hence could not be further improved. It is scored against accuracy metric.

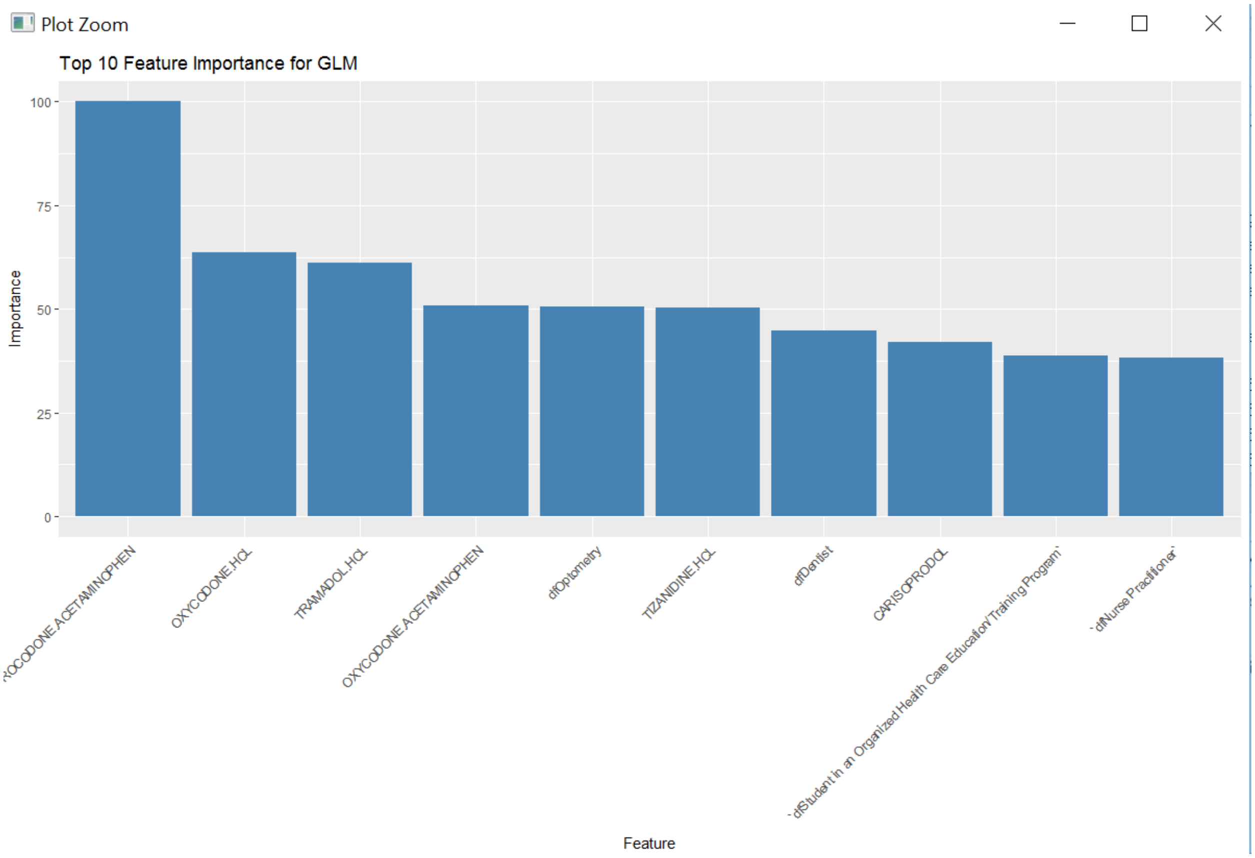


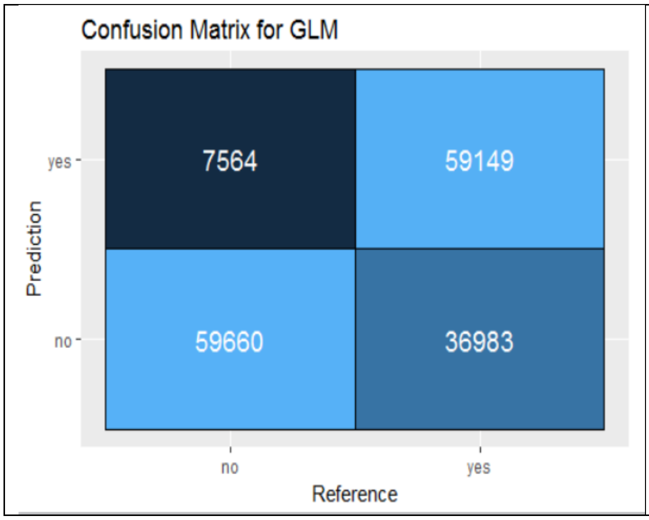



```
              Accuracy : 0.9222
                95% CI : (0.9209, 0
   No Information Rate : 0.5885
   P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 0.8409
Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.9104
           Specificity : 0.9390
        Pos Pred Value : 0.9552
        Neg Pred Value : 0.8799
            Prevalence : 0.5885
        Detection Rate : 0.5358
  Detection Prevalence : 0.5609
     Balanced Accuracy : 0.9247

      'Positive' Class : yes
```

LOGISTIC REGRESSION (GLM):

GLM is optimized using caret train control cross validation using 5 folds/no-repeat and binomial family. It is scored against accuracy metric. It is used for its ease of understanding.

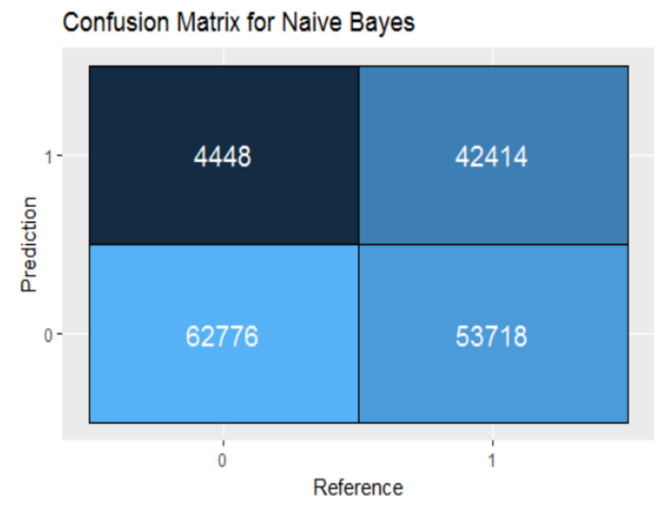


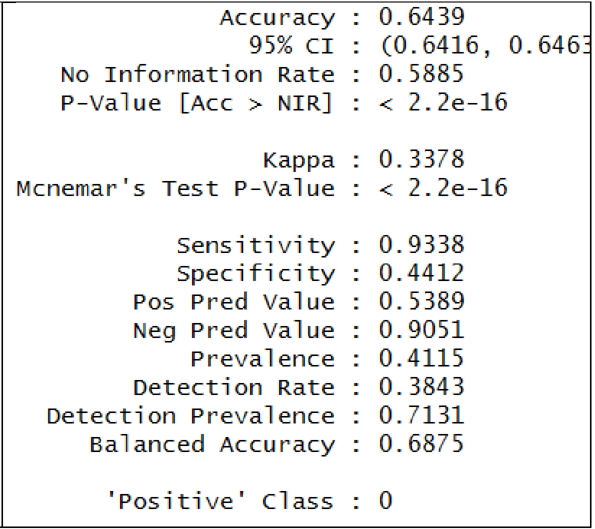


```
                Accuracy : 0.7273
                  95% CI : (0.7251, 0.7295)
     No Information Rate : 0.5885
     P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.4717
  Mcnemar's Test P-Value : < 2.2e-16

             Sensitivity : 0.6153
             Specificity : 0.8875
          Pos Pred Value : 0.8866
          Neg Pred Value : 0.6173
              Prevalence : 0.5885
          Detection Rate : 0.3621
    Detection Prevalence : 0.4084
       Balanced Accuracy : 0.7514

        'Positive' Class : yes
```

NAÏVE BAYES:

The reason for choosing Naïve Bayes is that it is easy to understand and does not take correlated drug columns into account. Laplace smoothing is not used and it performed very poorly.

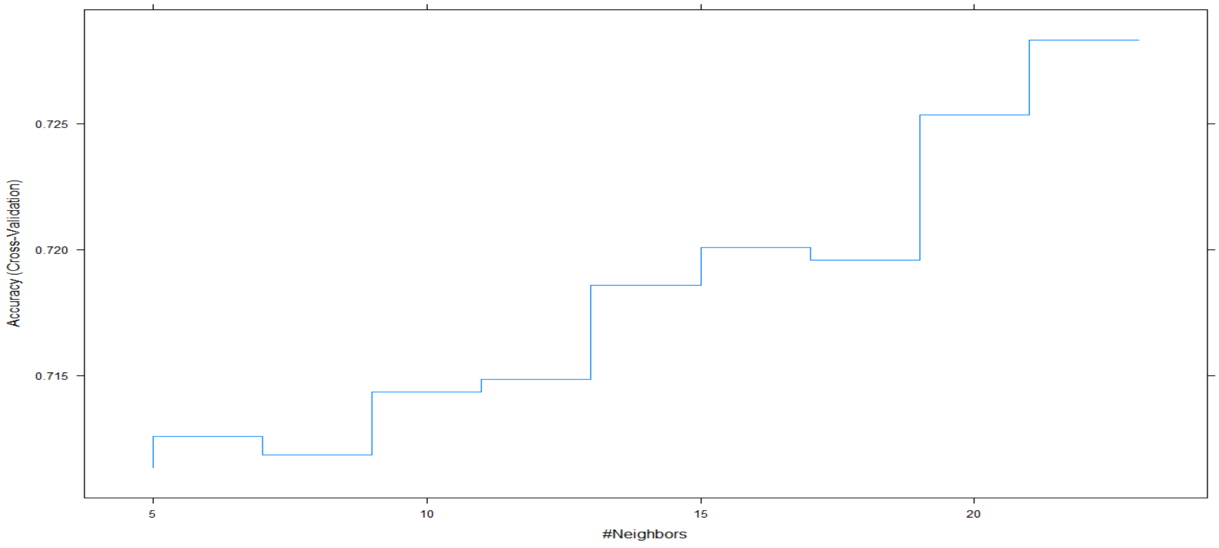


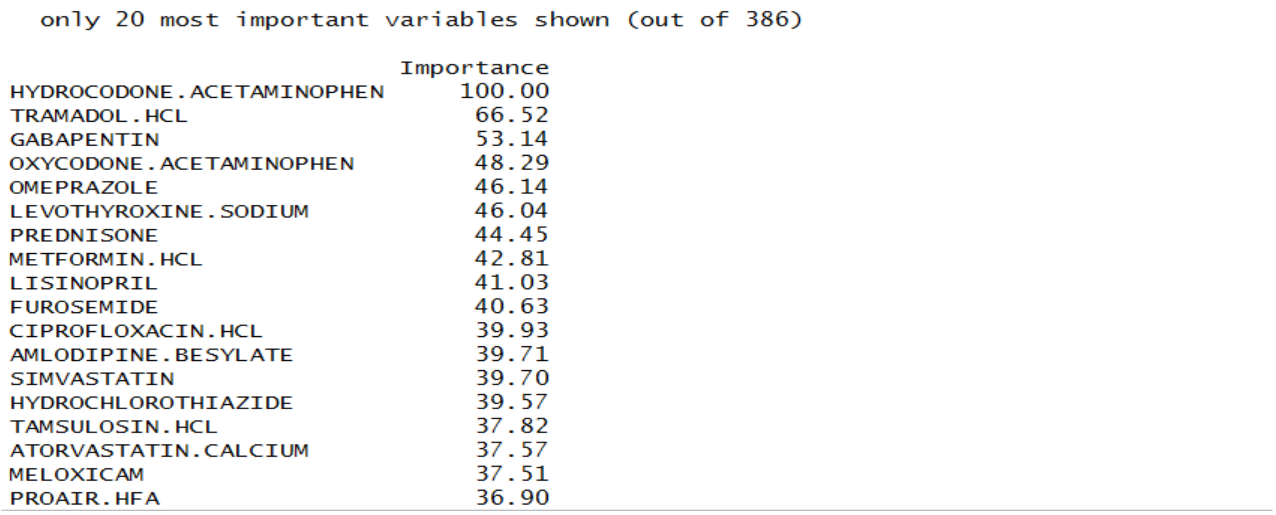


```
              Accuracy : 0.6439
                95% CI : (0.6416, 0.6463
   No Information Rate : 0.5885
   P-Value [Acc > NIR] : < 2.2e-16

                 Kappa : 0.3378
Mcnemar's Test P-Value : < 2.2e-16

           Sensitivity : 0.9338
           Specificity : 0.4412
        Pos Pred Value : 0.5389
        Neg Pred Value : 0.9051
            Prevalence : 0.4115
        Detection Rate : 0.3843
  Detection Prevalence : 0.7131
     Balanced Accuracy : 0.6875

      'Positive' Class : 0
```

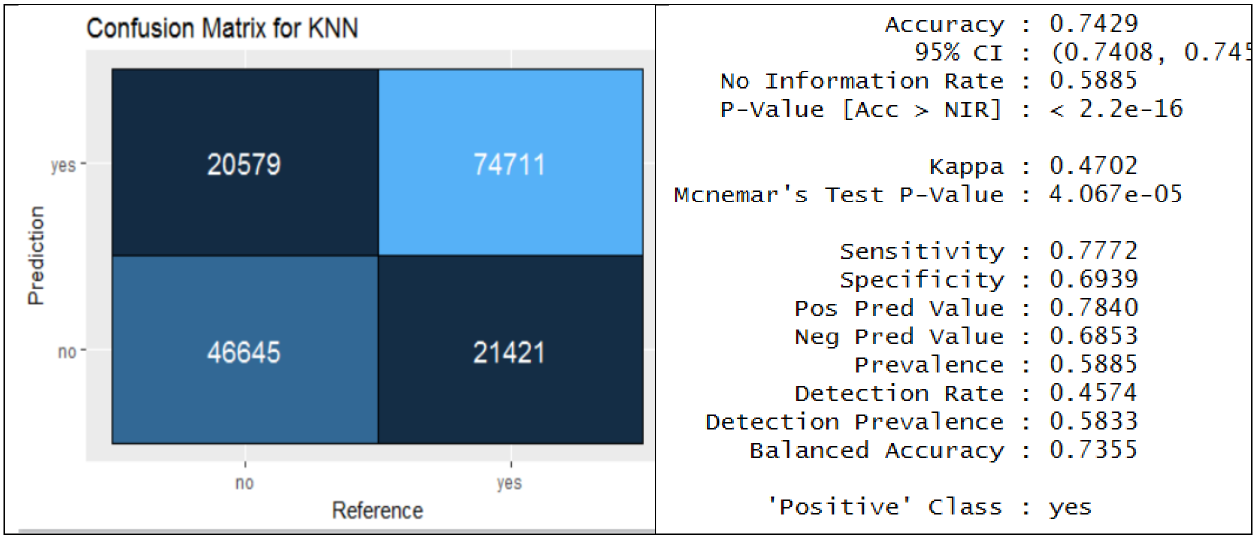
K-NEAREST NEIGHBOUR:

KNN is optimized using caret train control cross validation using 5 folds/no-repeat. It is scaled using "pre-Process" function in caret. It is scored against accuracy metric and best performed at k=23.





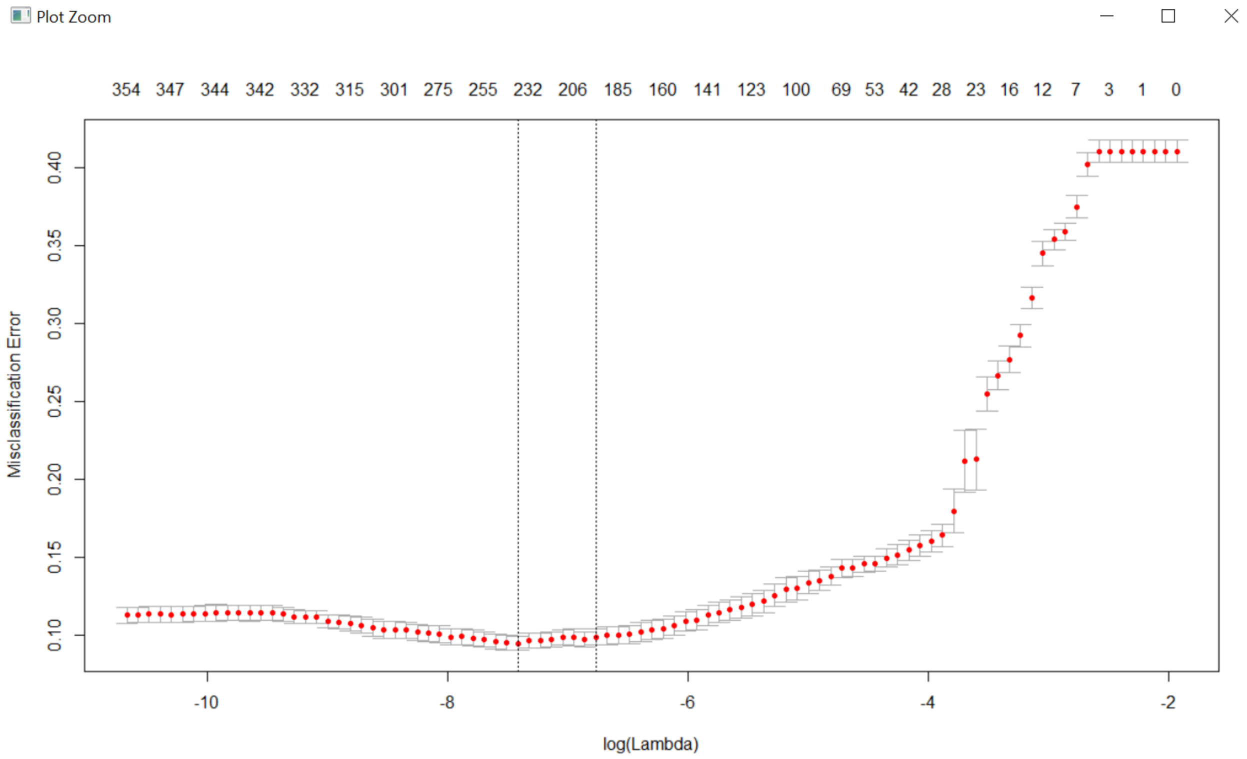
```
  only 20 most important variables shown (out of 386)

                             Importance
HYDROCODONE.ACETAMINOPHEN       100.00
TRAMADOL.HCL                     66.52
GABAPENTIN                       53.14
OXYCODONE.ACETAMINOPHEN          48.29
OMEPRAZOLE                       46.14
LEVOTHYROXINE.SODIUM             46.04
PREDNISONE                       44.45
METFORMIN.HCL                    42.81
LISINOPRIL                       41.03
FUROSEMIDE                       40.63
CIPROFLOXACIN.HCL                39.93
AMLODIPINE.BESYLATE              39.71
SIMVASTATIN                      39.70
HYDROCHLOROTHIAZIDE              39.57
TAMSULOSIN.HCL                   37.82
ATORVASTATIN.CALCIUM             37.57
MELOXICAM                        37.51
PROAIR.HFA                       36.90
```

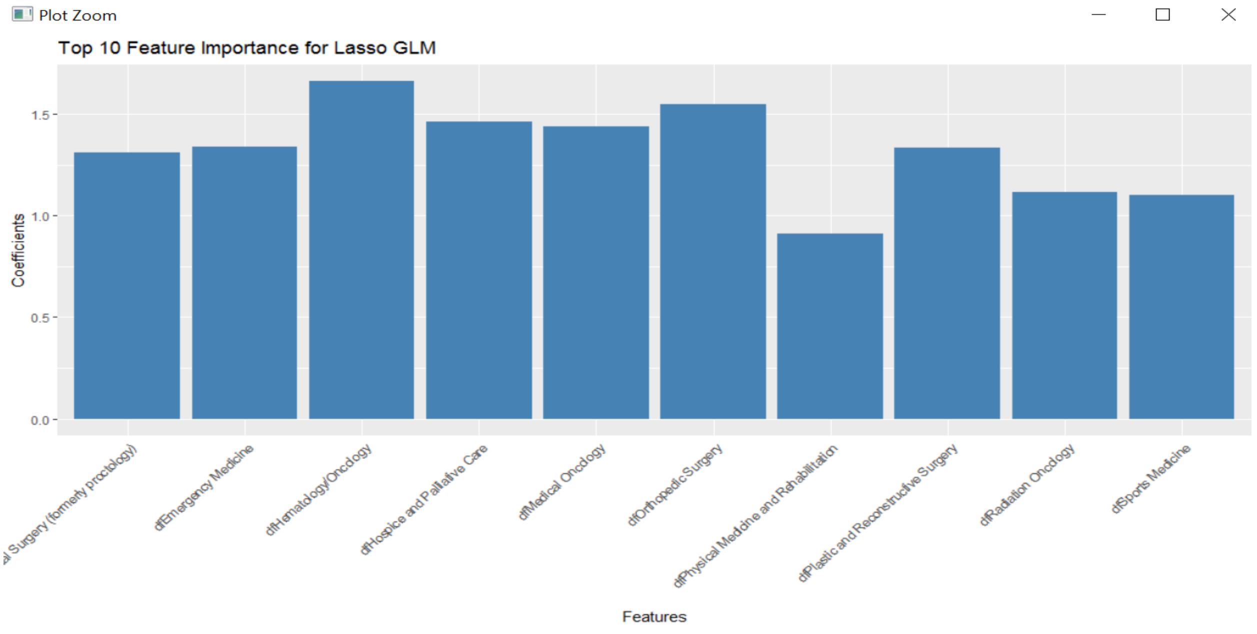


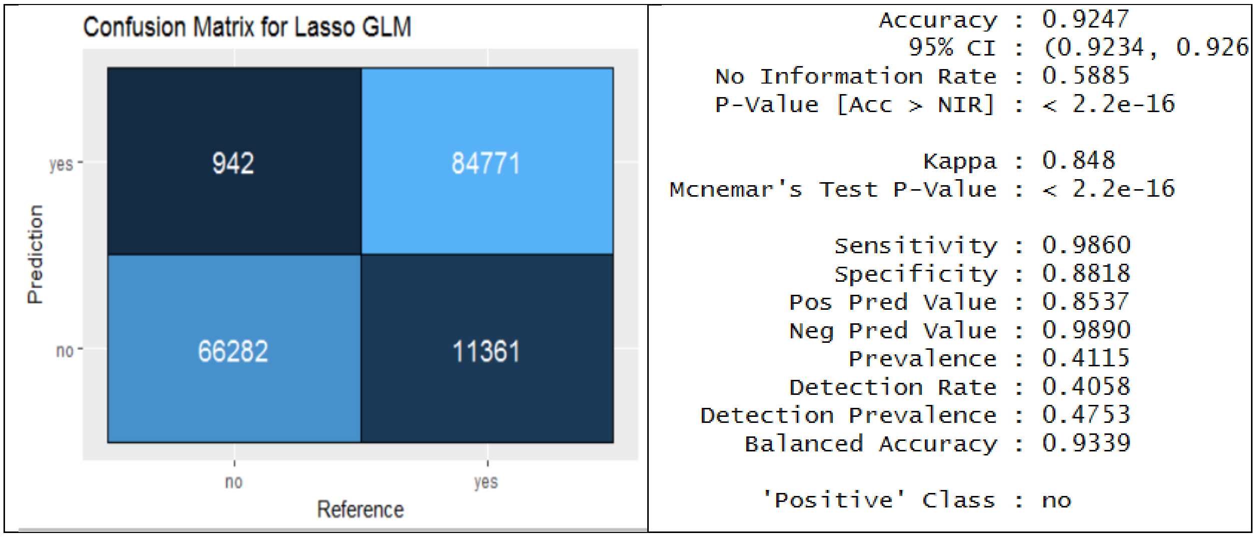
```
               Accuracy : 0.7429
                 95% CI : (0.7408, 0.74
    No Information Rate : 0.5885
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4702
 Mcnemar's Test P-Value : 4.067e-05

            Sensitivity : 0.7772
            Specificity : 0.6939
         Pos Pred Value : 0.7840
         Neg Pred Value : 0.6853
             Prevalence : 0.5885
         Detection Rate : 0.4574
   Detection Prevalence : 0.5833
      Balanced Accuracy : 0.7355

       'Positive' Class : yes
```

LOGISTIC REGRESSION (GLM) WITH LASSO:

Lasso is optimized using caret train control cross validation using 5 folds/no-repeat to find the optimal lambda value. It is scored against accuracy metric. It is great for identifying features but CV train time is insanely high.


Plot Zoom
354 347 344 342 332 315 301 275 255 232 206 185 160 141 123 100 69 53 42 28 23 16 12 7 3 1 0
Misclassification Error
0.10
0.15
0.20
0.25
0.30
0.35
0.40
-10
-8
-6
-4
-2
log(Lambda)

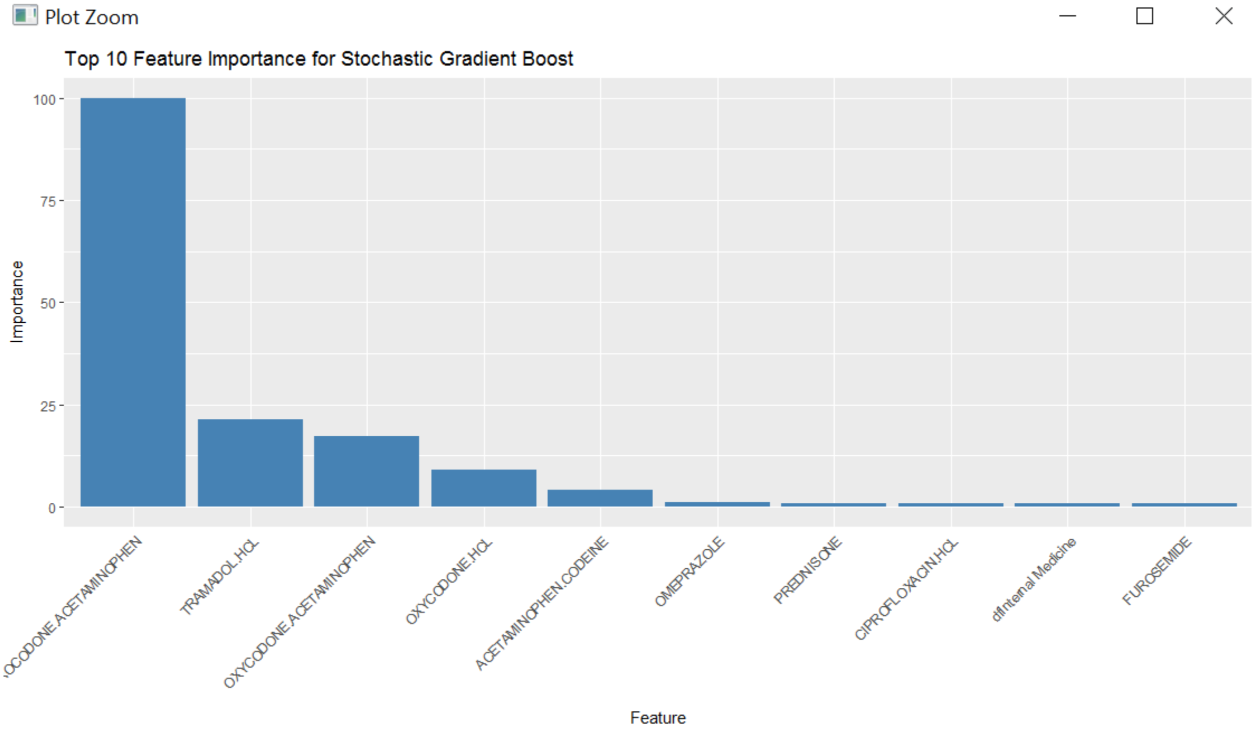

Plot Zoom
Top 10 Feature Importance for Lasso GLM
Coefficients
0.0
0.5
1.0
1.5
al Surgery (formerly proctology)
dfEmergency Medicine
dfHematology/Oncology
dfHospice and Palliative Care
dfMedical Oncology
dfOrthopedic Surgery
dfPhysical Medicine and Rehabilitation
dfPlastic and Reconstructive Surgery
dfRadiation Oncology
dfSports Medicine
Features

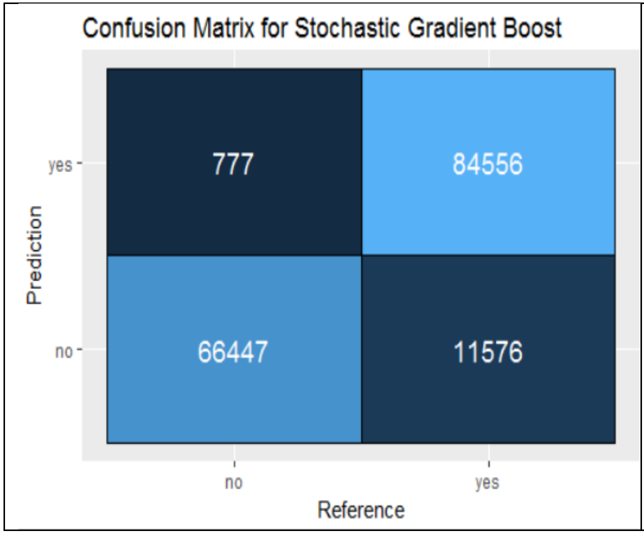


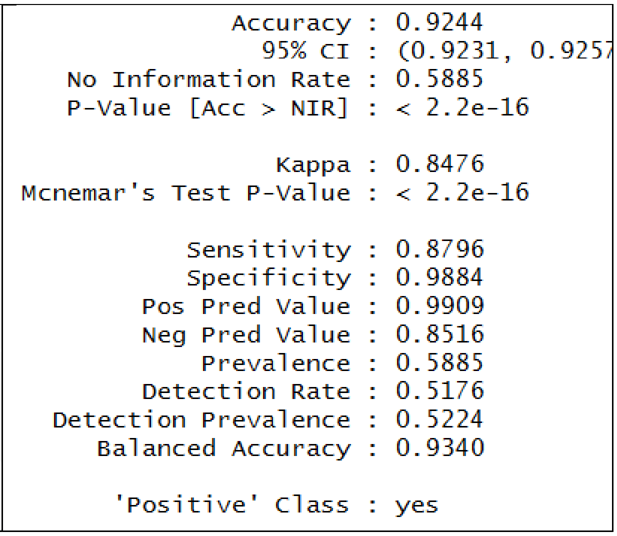
```
                  Accuracy : 0.9247
                    95% CI : (0.9234, 0.926
       No Information Rate : 0.5885
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.848
    Mcnemar's Test P-Value : < 2.2e-16

               Sensitivity : 0.9860
               Specificity : 0.8818
            Pos Pred Value : 0.8537
            Neg Pred Value : 0.9890
                Prevalence : 0.4115
            Detection Rate : 0.4058
      Detection Prevalence : 0.4753
         Balanced Accuracy : 0.9339

          'Positive' Class : no
```

STOCHASTIC GRADIENT BOOSTING:

GBM is optimized using caret train control cross validation using 5 folds/no-repeat. It trained very fast and gave a good accuracy.





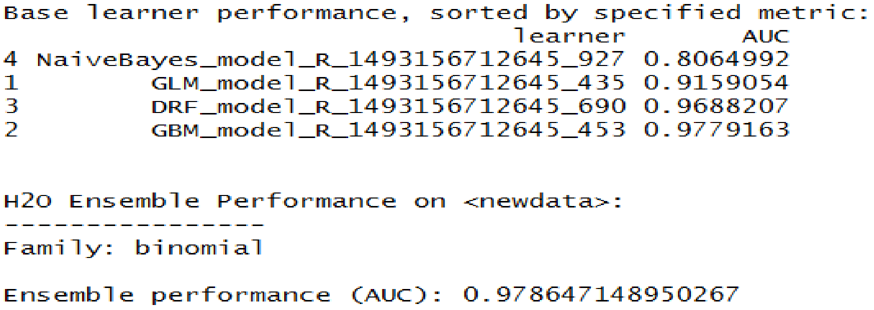


```
          Accuracy : 0.9244
            95% CI : (0.9231, 0.9257
No Information Rate : 0.5885
P-Value [Acc > NIR] : < 2.2e-16

             Kappa : 0.8476
Mcnemar's Test P-Value : < 2.2e-16

       Sensitivity : 0.8796
       Specificity : 0.9884
    Pos Pred Value : 0.9909
    Neg Pred Value : 0.8516
        Prevalence : 0.5885
    Detection Rate : 0.5176
Detection Prevalence : 0.5224
  Balanced Accuracy : 0.9340

   'Positive' Class : yes
```

ENSEMBLE TECHNIQUE:

Models are ensemble to give a higher accuracy.

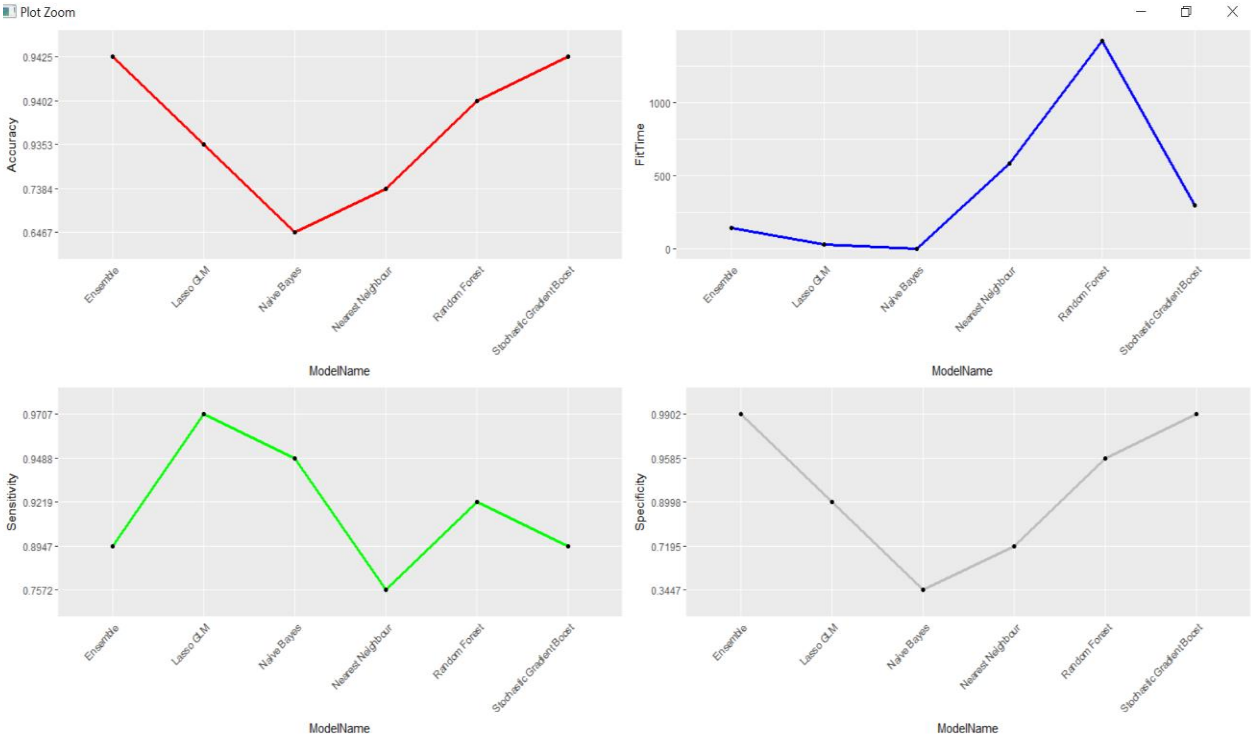
- H2o package is used for stacking.
- Performance tuned parameters from standalone models are used in stacking instead of cross validating again.
- Base model wrappers are created for each models and parameters
- Meta learner used is logistic regression
- Base models are passed as a list to stack on meta learner.
- Model performance measured was better than standalone models.



```
Base learner performance, sorted by specified metric:
                              learner       AUC
4  NaiveBayes_model_R_1493156712645_927 0.8064992
1         GLM_model_R_1493156712645_435 0.9159054
3         DRF_model_R_1493156712645_690 0.9688207
2         GBM_model_R_1493156712645_453 0.9779163


H2O Ensemble Performance on <newdata>:
----------------
Family: binomial

Ensemble performance (AUC): 0.978647148950267
```

MODEL PERFORMANCE COMPARISON

The figure shows comparative model performance of base models and ensemble. Considering the accuracy and model fit time, ensemble and stochastic gradient boost is a winner.



ISSUES

Data Set:

- Features and the data size along with hardware limitations increased the runtime to days.
- Data set required a lot of cleaning to bring it to a suitable format.
- Merging and Data validation was cumbersome.
- Cross validation time on some models was very high even with 5 folds and no repeats.
- Memory issues.

Collinearity:

Due to presence of feature “opioid_claims/refills” all models gave 100% accuracy. Statistically this model was highly important to the target label with zero p-value. However, the reason is unknown as the collinearity could not be established because target label is binary. The models performed realistically after removing this column. More investigation into it is needed.

SUMMARY

- The consumption of opioids is higher amongst white race across all age groups.
- Compounded drugs are to be watched out for mostly ACETAMENOPHEN and HCL.
- HYDROCODONE.ACETAMENOPHEN is a non-opioid drug of paramount importance.
- Female doctors are more inclined to prescribing opiates.
- The U.S. boundary states look more prone to opioid overdose deaths.
- The ensemble model is good predictor of opioid prescriptions with 94.4% accuracy and AUC 97.2%. Stochastic gradient boost works best as a base model.

NEXT STEP & FUTURE WORK

- High number of refills must be related to higher overdose deaths as well as addictions. But this could not be linked to the demographic data. Further investigation is needed.
- There is a relation between opioid deaths and the prescriber's speciality. Some of the selected specialities are counter intuitive. Domain knowledge will be helpful.
- Model performance can be further improved. Zero variance columns could have been removed beforehand and sub-setting to top 250 drugs could have been expanded.

REFERENCES

- Centre of Medicare and Medicaid services - https://www.cms.gov/
- Centre for disease control and prevention - https://wonder.cdc.gov/controller/datarequest/D76
- Library documentations - dplyr, ggplot, h2o, caret
- Issue resolution and debugging - Rbloggers, Kaggle