Neil Bhutada

Nacewicz Lab

Independent Research Paper

Words: 683

Interesting parts: 2nd part of the 5th page – till the end

Using Gradient Boosting Machines to understand depression

To understand what Gradient Boosting Machine (GBM) is, we need to look at what Decision Trees and Random Forests are. A Decision tree is a representation of logical/structured human thinking demonstrated by trees in Computer Science. Each node represents a condition, and each edge represents whether the condition is true or false – for an example look at fig. 1. Now, a Random forest is a culmination of various decision trees where each tree is trained on a random subset(s) – where a few trees would use certain rows and columns only of data. Each tree is built independently, and the result of the model will be average/majority of what each individual tree has produced (Stephanie Glen). On the other hand, a GBM follows more of an additive process where each tree is built one at a time, and a new tree is built from its previous one. The weaker rows (or types of rows that have been predicted inaccurately) are 'boosted' by adding some weights to them. Due to the process followed by GBM, it is excellent at creating models from Datasets that are non-linear/erratic. However, the con of using GBMs is that if it is not used carefully, they can overfit by capturing the noise of the dataset as well.

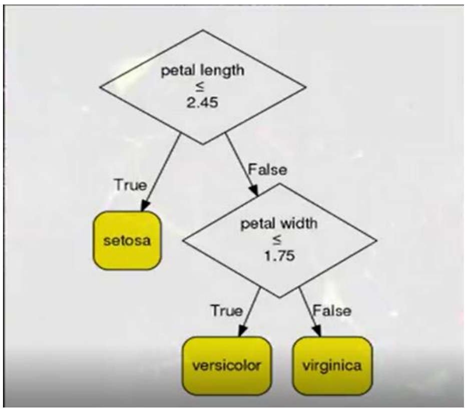
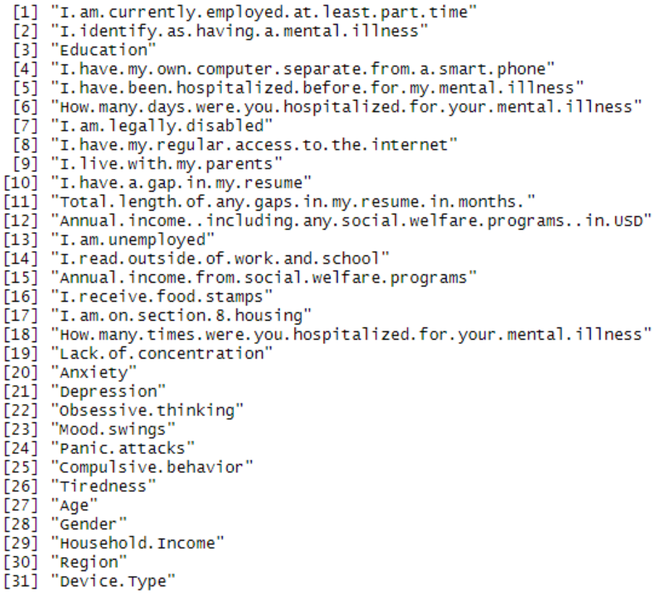


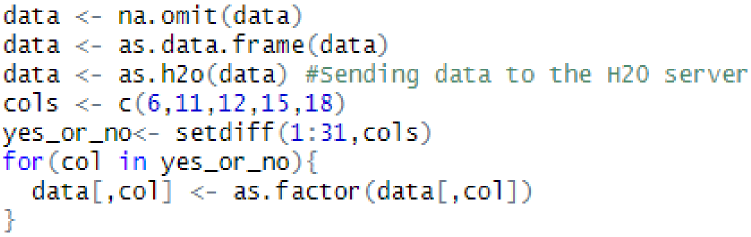
Figure 1

In this paper, we would be seeing how we could use GBM to determine whether a person has depression based on his employment status, prior mental illnesses, gadgets used by the person, personal details (such as gender, age, location), and hospital admits. The dataset I would using is by Michael Corley (the link could be found in the references section). To obtain the best accuracy for this dataset, a GBM was used. The GBM could find suitable patterns between non-linear attributes such as the relationship between age groups and depression. The dataset has 31 columns and 334 rows (refer to figure 2 for column names). The only pre-processing that had to be done was changing the column types with values of 0's and 1's from integer to factor columns; this was done to make sure that the GBM does not perform a regression, and performs a classification (to be more specific logistic regression) instead. Rows with value 'na' were dropped. The code for pre-processing the data could be found in figure 3.



```
 [1] "I.am.currently.employed.at.least.part.time"
 [2] "I.identify.as.having.a.mental.illness"
 [3] "Education"
 [4] "I.have.my.own.computer.separate.from.a.smart.phone"
 [5] "I.have.been.hospitalized.before.for.my.mental.illness"
 [6] "How.many.days.were.you.hospitalized.for.your.mental.illness"
 [7] "I.am.legally.disabled"
 [8] "I.have.my.regular.access.to.the.internet"
 [9] "I.live.with.my.parents"
[10] "I.have.a.gap.in.my.resume"
[11] "Total.length.of.any.gaps.in.my.resume.in.months."
[12] "Annual.income..including.any.social.welfare.programs..in.USD"
[13] "I.am.unemployed"
[14] "I.read.outside.of.work.and.school"
[15] "Annual.income.from.social.welfare.programs"
[16] "I.receive.food.stamps"
[17] "I.am.on.section.8.housing"
[18] "How.many.times.were.you.hospitalized.for.your.mental.illness"
[19] "Lack.of.concentration"
[20] "Anxiety"
[21] "Depression"
[22] "Obsessive.thinking"
[23] "Mood.swings"
[24] "Panic.attacks"
[25] "Compulsive.behavior"
[26] "Tiredness"
[27] "Age"
[28] "Gender"
[29] "Household.Income"
[30] "Region"
[31] "Device.Type"
```

Figure 2



```
data <- na.omit(data)
data <- as.data.frame(data)
data <- as.h2o(data) #Sending data to the H20 server
cols <- c(6,11,12,15,18)
yes_or_no<- setdiff(1:31,cols)
for(col in yes_or_no){
  data[,col] <- as.factor(data[,col])
}
```

Figure 3

After pre-processing the data, a GBM model was created. The GBM model used cross validation to prevent over-fitting. The number of folds for cross-validation used was 5, and the fold types was ‘modulo’ (which allows even splits of the data sets). Early stopping was also used to make sure that overfitting does not take place. The number of trees used were 40, with the maximum depth being 8 and minimum rows (fewest allowed weighted leaf nodes) being 15. AutoML was used as an inspiration for setting the model parameters. Figures for model training metrics could be found in figures 4,5,6. Fig 7 includes result on test data.

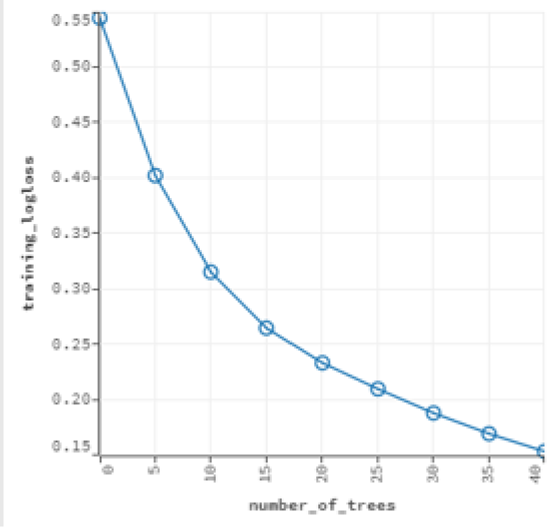


Figure 4

This is the relationship between number of trees used and the log loss on the training data obtained after each tree.

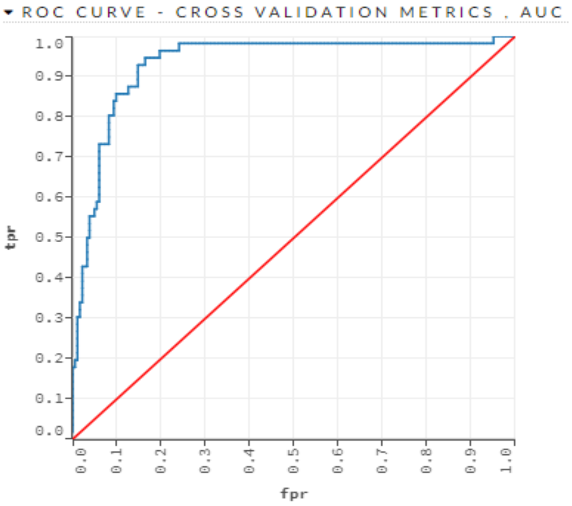
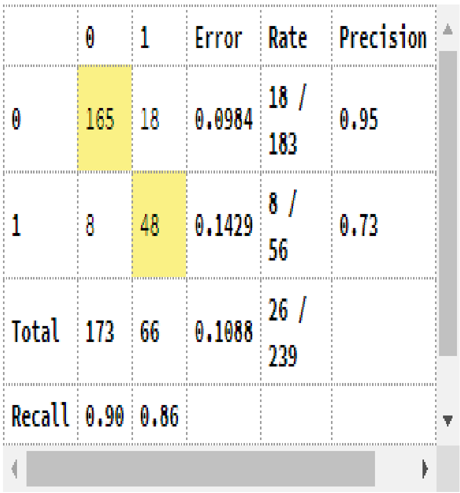


Figure 5

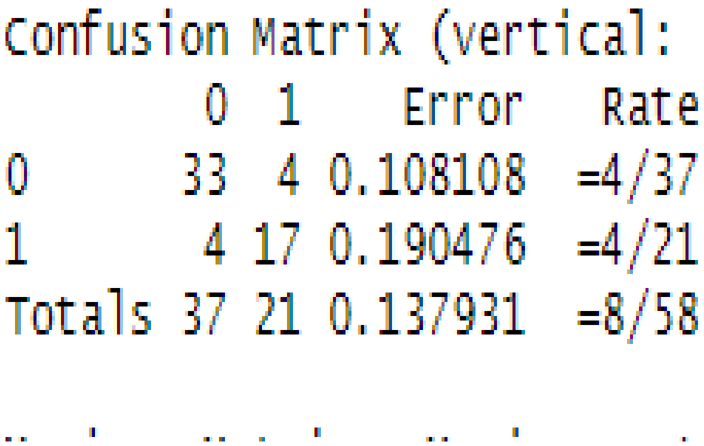
This is the ROC curve which is plotted as the False Positive Rate vs True Positive Rate. The Area Under the Curve (AUC) is near 1.0 (which is good).



	0	1	Error	Rate	Precision
0	165	18	0.0984	18 / 183	0.95
1	8	48	0.1429	8 / 56	0.73
Total	173	66	0.1088	26 / 239	
Recall	0.90	0.86			

Figure 6

This is the confusion matrix on the validation dataset. The error of the model was around 10%.



```
Confusion Matrix (vertical:
          0  1     Error   Rate
0        33  4 0.108108  =4/37
1         4 17 0.190476  =4/21
Totals   37 21 0.137931  =8/58
```

Figure 7

This is the confusion matrix on the test dataset. The error of the model was around 13%.

Now coming to the analysis of determining whether a person is depressed by the GBM, the most important medical factor was whether the person identified as having a mental illness; according the most important non- medical factor was the person's house hold income. (Look at figure 8 for more details)

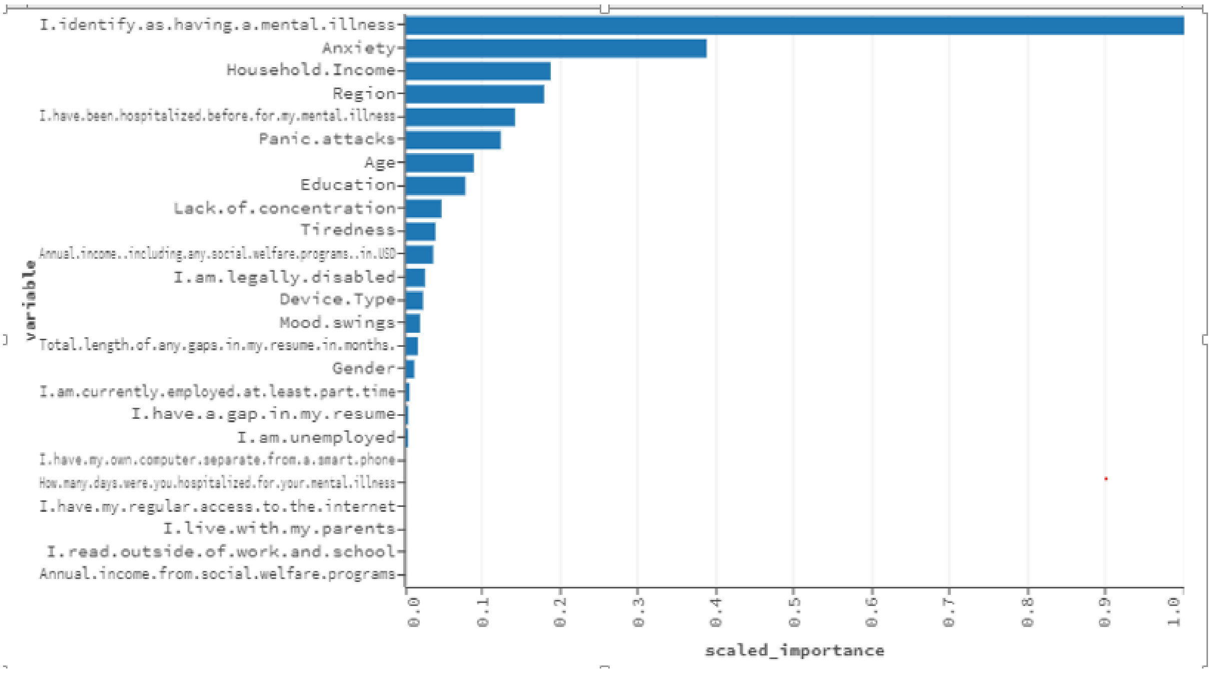
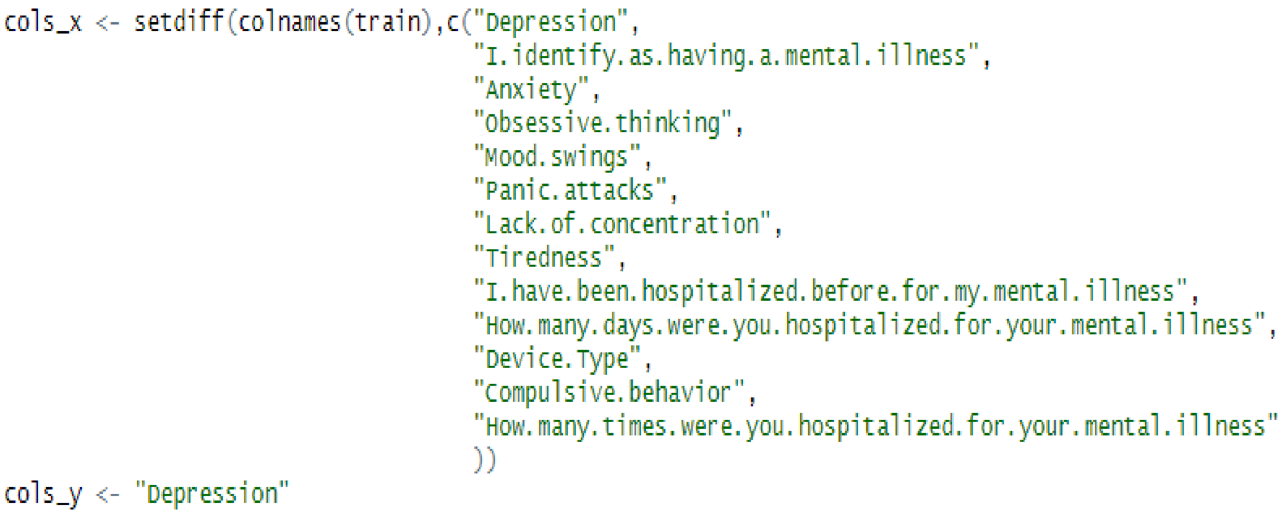


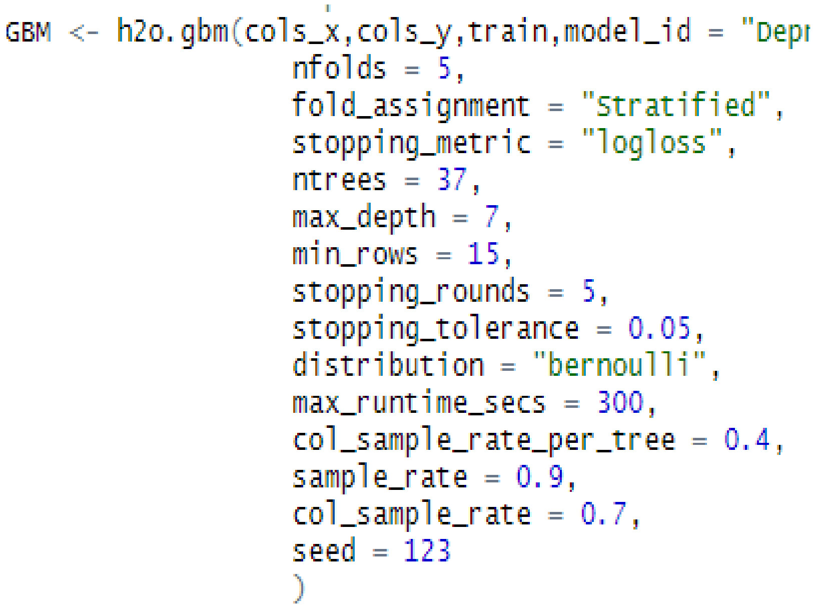
Figure 8

There is a positive co-relation between the levels and values.

But wait! This is common sense, right? It does kind of seem obvious that if a person has identified themselves as having a mental illness, has anxiety or any other medical factors would most likely have depression. Also, we haven't taken the full advantage of GBMs, which is the ability to predict values on linear data. Therefore, I made another GBM and excluded all medical related columns in the training data.



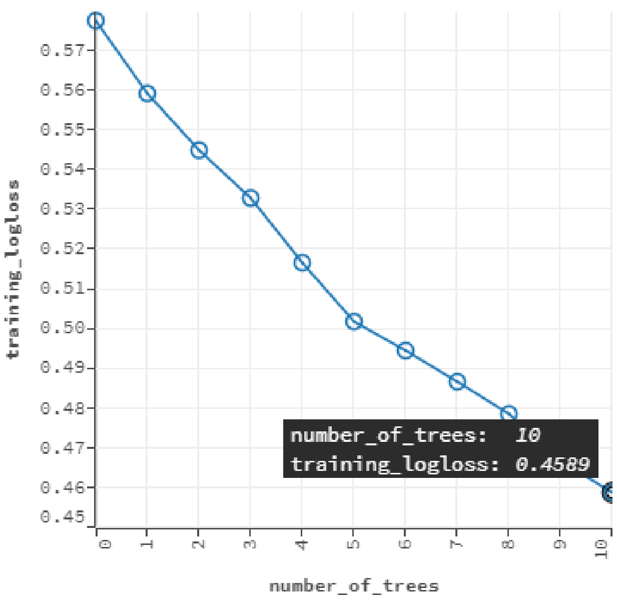
```
cols_x <- setdiff(colnames(train),c("Depression",
                                     "I.identify.as.having.a.mental.illness",
                                     "Anxiety",
                                     "Obsessive.thinking",
                                     "Mood.swings",
                                     "Panic.attacks",
                                     "Lack.of.concentration",
                                     "Tiredness",
                                     "I.have.been.hospitalized.before.for.my.mental.illness",
                                     "How.many.days.were.you.hospitalized.for.your.mental.illness",
                                     "Device.Type",
                                     "Compulsive.behavior",
                                     "How.many.times.were.you.hospitalized.for.your.mental.illness"
                                     ))
cols_y <- "Depression"
```



```
GBM <- h2o.gbm(cols_x,cols_y,train,model_id = "Depr
               nfolds = 5,
               fold_assignment = "Stratified",
               stopping_metric = "logloss",
               ntrees = 37,
               max_depth = 7,
               min_rows = 15,
               stopping_rounds = 5,
               stopping_tolerance = 0.05,
               distribution = "bernoulli",
               max_runtime_secs = 300,
               col_sample_rate_per_tree = 0.4,
               sample_rate = 0.9,
               col_sample_rate = 0.7,
               seed = 123
               )
```

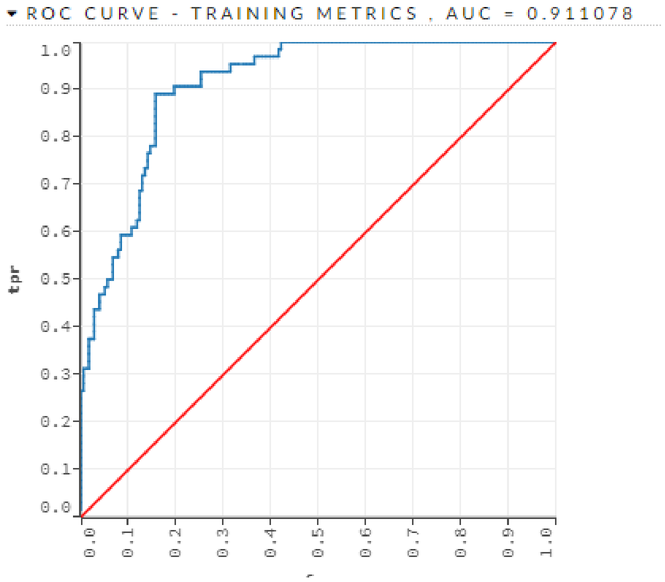
In the parameter of the GBM, fold assignment is stratified because these allows even splits of data during cross-validation such that even the various levels of data are evenly split. This split enables the cross-validation prediction to be more uniform, and thus trains the model a better way. Log loss is used as an error measuring metric because this is logical regression. Stopping rounds , tolerance, and sample rates of cols and rows prevent overfitting. Stopping rounds, stops training the model if after a certain number rounds the model's error is not lower than the stopping tolerance. Sample rate, randomly choose a certain number of rows/columns to be train by the GBM in each iteration or each tree. Maybe, for GBMs, iterations is equal to the number of nfolds. The distribution used by the model is Bernoulli.

After running the model, we get the following results:



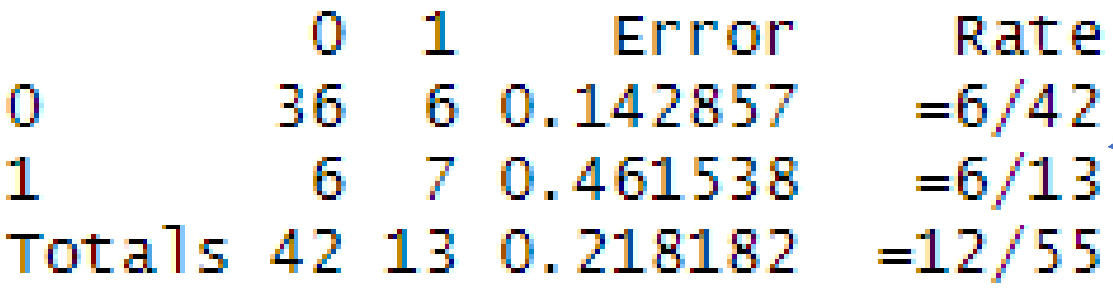
This is the scoring history of the model.

Most likely, the model stopped training after 10 trees to prevent over-fitting.



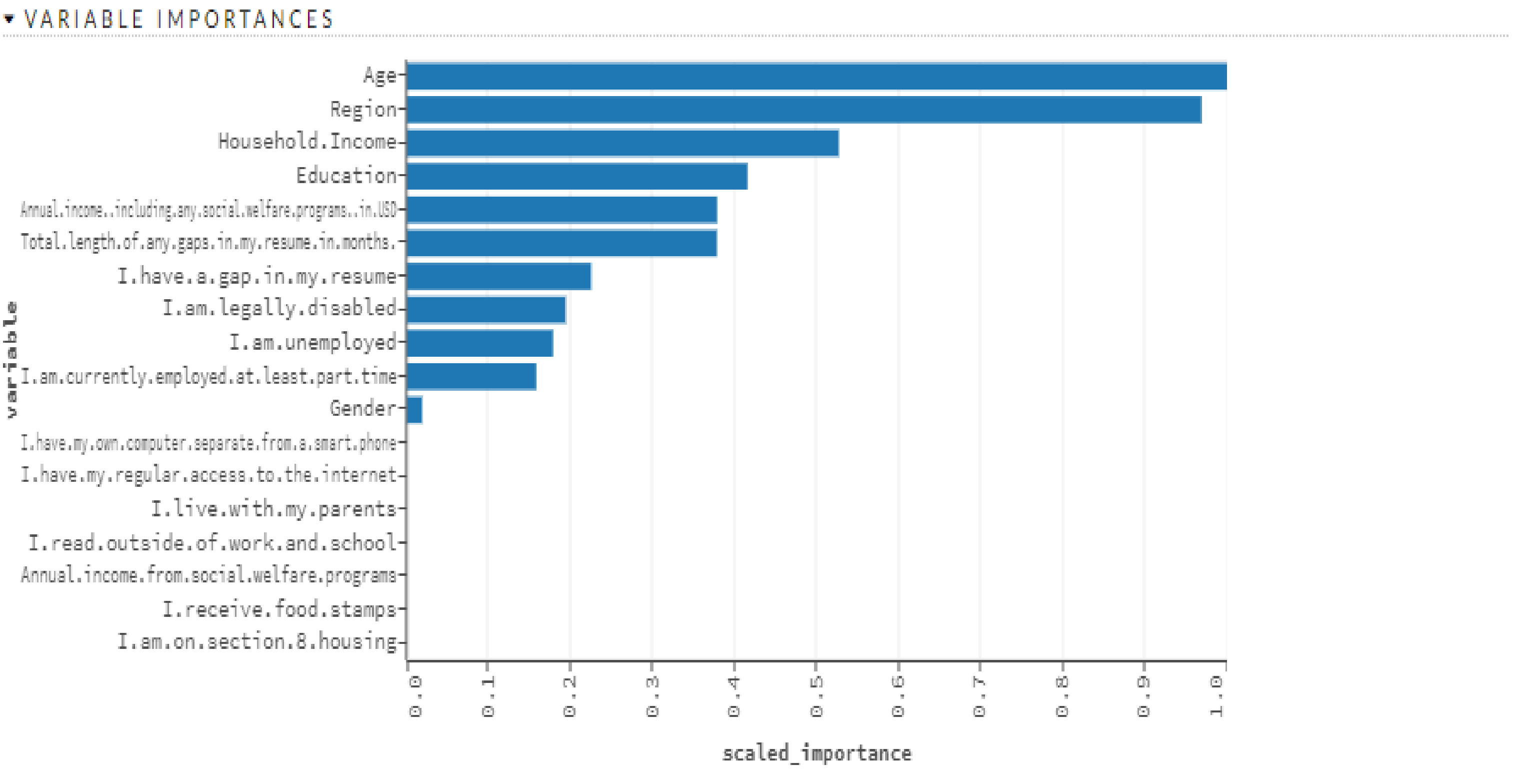
This is the ROC curve plotted against false positive rate (total number of correct 0's by model/total number of zero's) vs true positive rate (total number of correct 1's by model/total number of 1's)

The AUC of this curve is 0.91 which is close to 1. This implies the model has trained well.

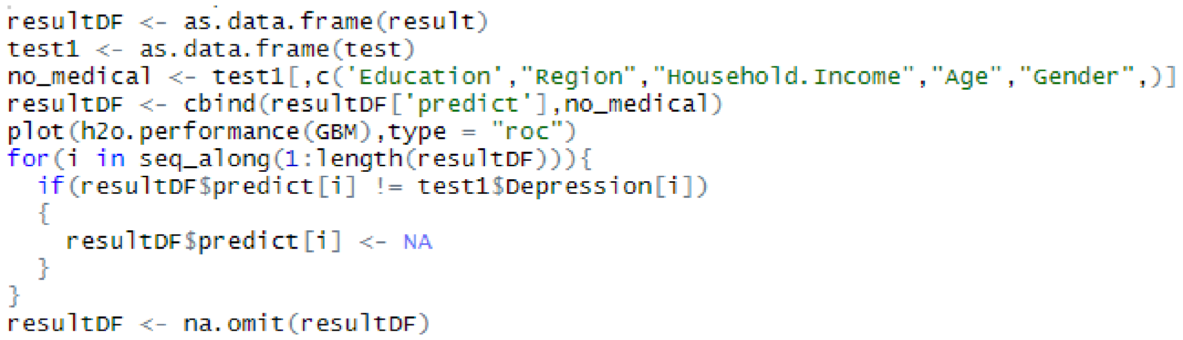


```
         0  1    Error    Rate
0       36  6 0.142857  =6/42
1        6  7 0.461538  =6/13
Totals  42 13 0.218182 =12/55
```

Confusion matrix on test data set.

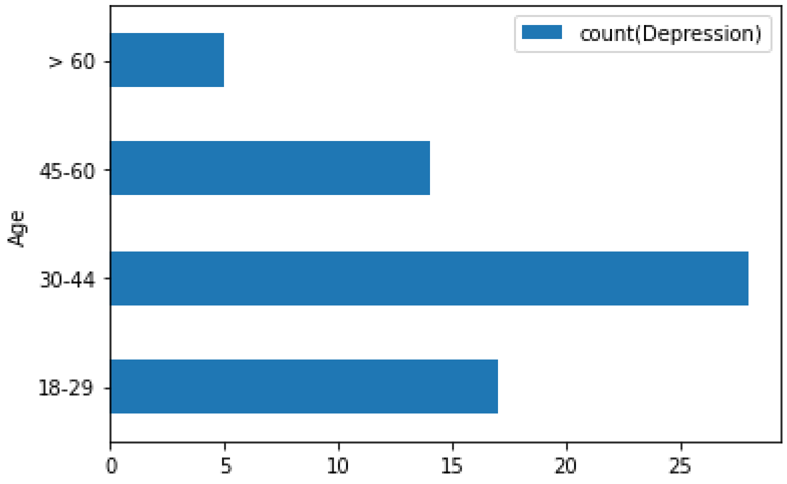


Now in the variable importance graph, it says that all the columns are positively co-related to depression. However, the graph does not specify how various factors in a column relate to depression. Therefore, to do so, I plotted a graphs from the training data for the important columns against depression, and I compared them to only the correct predictions the model made on the test data.

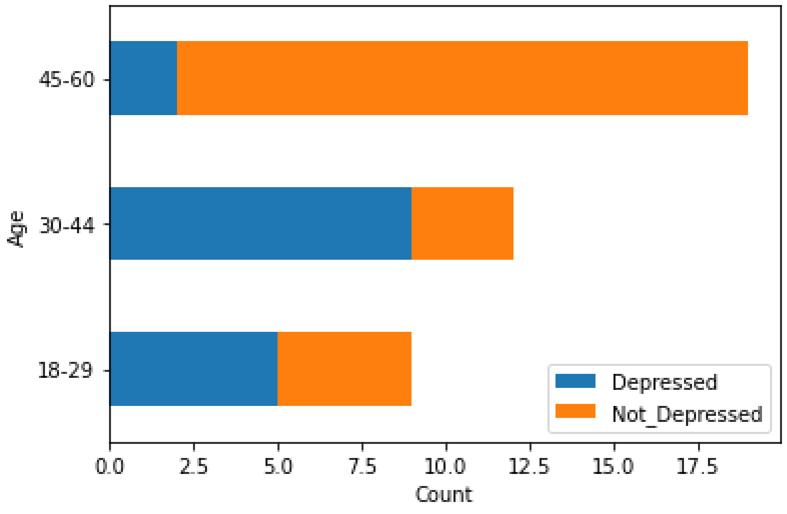


```
resultDF <- as.data.frame(result)
test1 <- as.data.frame(test)
no_medical <- test1[,c('Education',"Region","Household.Income","Age","Gender",)]
resultDF <- cbind(resultDF['predict'],no_medical)
plot(h2o.performance(GBM),type = "roc")
for(i in seq_along(1:length(resultDF))){
  if(resultDF$predict[i] != test1$Depression[i])
  {
    resultDF$predict[i] <- NA
  }
}
resultDF <- na.omit(resultDF)
```

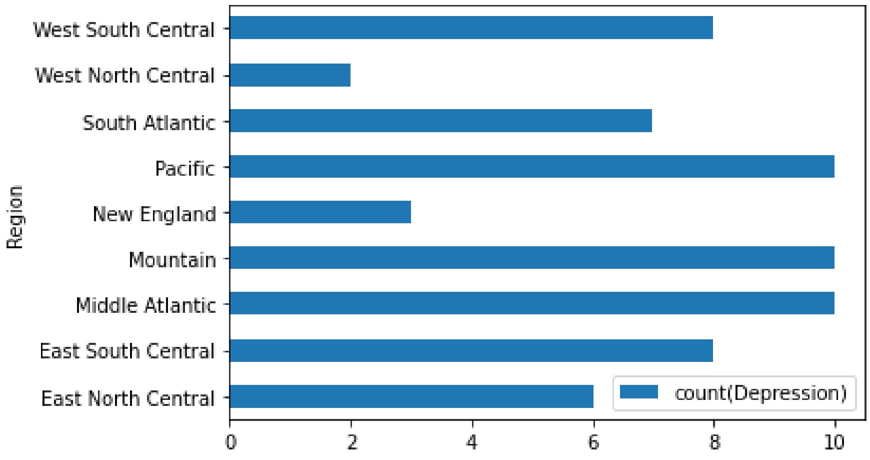
This is the code to make sure to extract only the accurate predictions from the test data in data frame called resultDF.



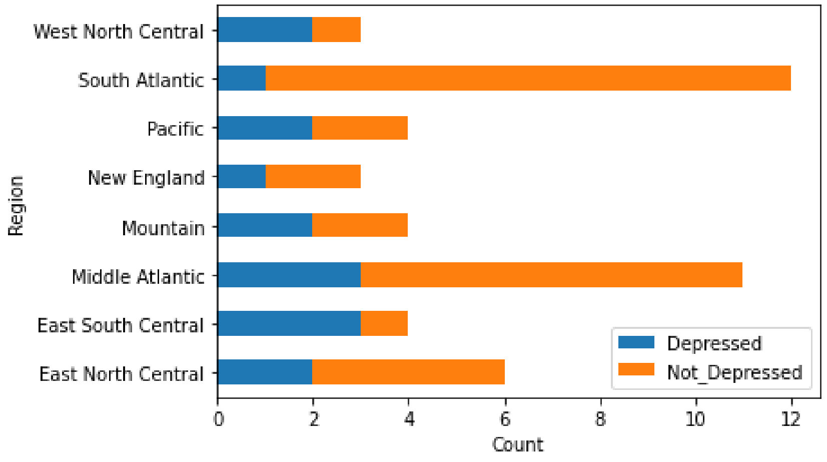
This is the graph of the different age groups against the number of depression cases from the training data set.



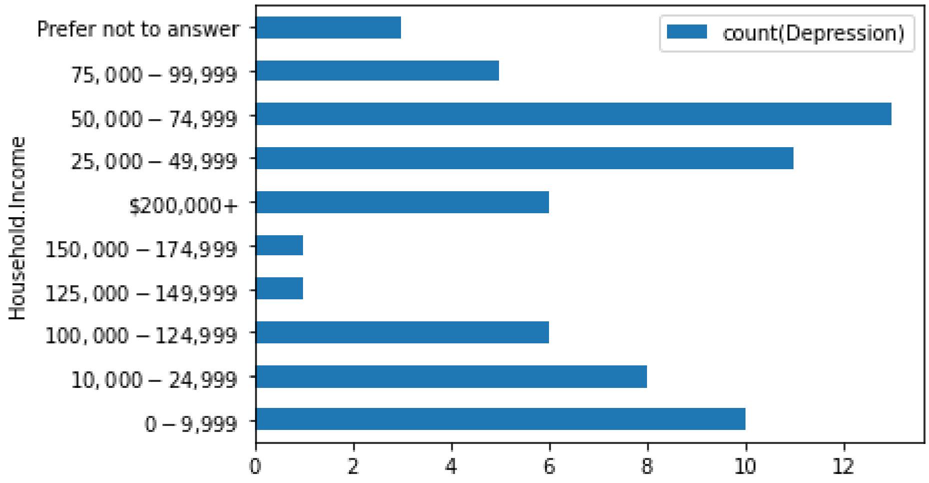
According to graph plotted from subset of the test dataset that contained only the correct predictions the model made, it is clearly visible the people aged from 30-44 were more likely depressed. This is because of the ratio of depressed to non-depressed.



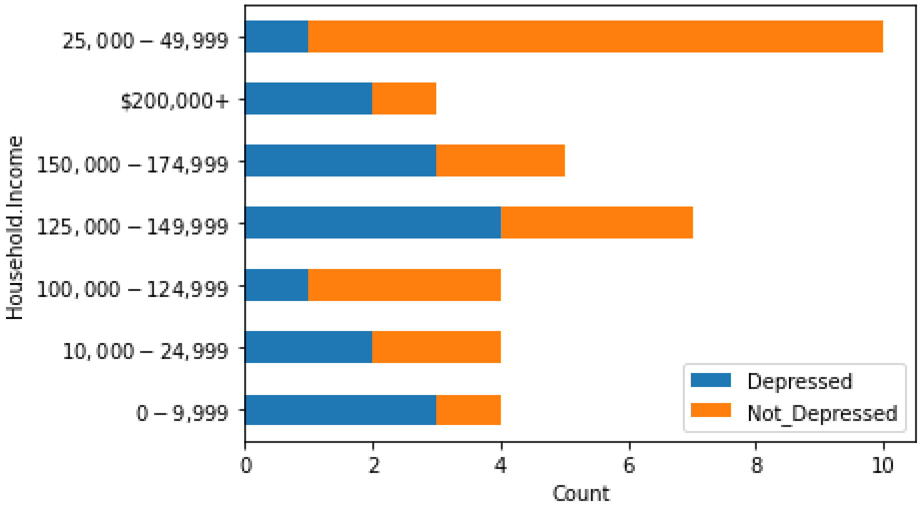
This is the graph of the different region groups against the number of depression cases from the training data set.



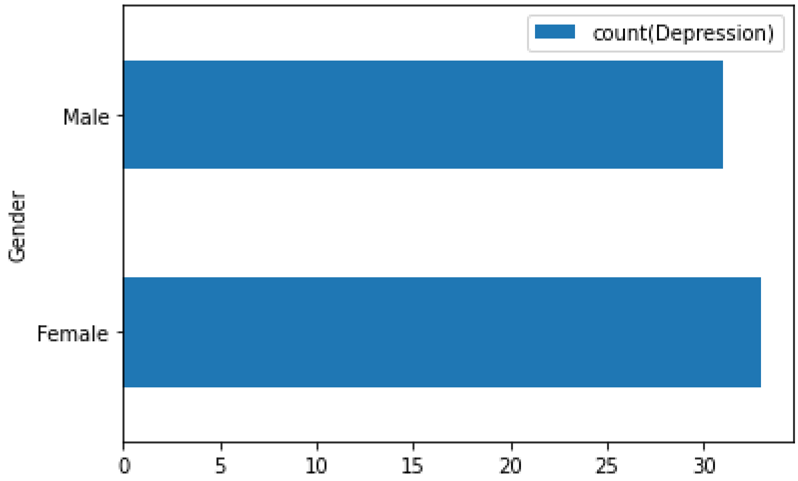
According to the test data set with only accurate predictions made by the model, people from West North Central and East South Central were more likely to be depressed. This is because of the ratio of depressed to non-depressed.



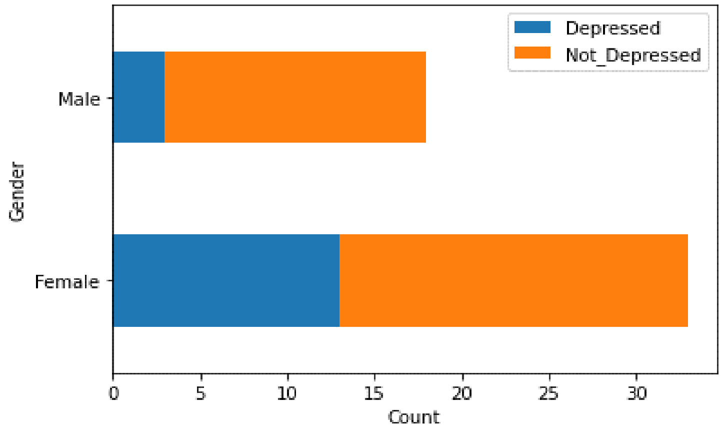
This is the graph of the different income groups against the number of depression cases from the training data set.



From this graph it could be seen that people with extreme income ends were more likely to be depressed. This is because of the ratio of depressed to non-depressed.



This is the graph of the different gender groups against the number of depression cases from the training data set.



According to the test data set with only accurate predictions made by the model, women were likely to be depressed compared to men. This is because of the ratio of depressed to non-depressed.

In this paragraph, we will compare our model's prediction to real world claims and evidence. Angst in her research paper stated that women are more likely to be depressed (Angst, 2020, para 1). According to a study conducted by Sareen et. al, low income groups were extremely likely to experience depression and suicidal thought (Sareen, 2011, Result's section). According to Assari et. al, usually high income was inversely proportional to likelihood of being depressed (Assari, 2017, Abstract section). However, in our dataset, 200k+ income ranged people were more likely to face depression. According to Mirowsky et. al, middle aged people were supposed to face the least amount of depression (Mirowsky et. al, 1992, 202). However, our model stated that people aged from 30 – 44 would be more likely depressed. Nevertheless, the paper written by Mirowski was written back in 1992 when the stress levels were way lower than the stress levels faced by people living in this time, and more people in their middle age during this generating deal with stress – a factor closely related to depression (Healthline.com). According to our data set, people from west north central and east south central should be at a higher risk of being depressed. Unfortunately, I could not find a source for specifically comparing depression and various regions in America; however, I found a source (mhanational.org) which talks about mental illnesses and regions in America. The west north central and east south-central regions did seem to be regions with higher mental illness cases.

References

Data set: https://www.kaggle.com/michaelacorley/unemployment-and-mental-illness-survey

Posted by Stephanie Glen on July 28, 2. (n.d.). Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply. Retrieved December 04, 2020, from https://www.datasciencecentral.com/profiles/blogs/decision-tree-vs-random-forest-vs-boosted-trees-explained

Angst, J., Gamma, A., Gastpar, M., L�Pine, J., Mendlewicz, J., & Tylee, A. (2002). Gender differences in depression. *European Archives of Psychiatry and Clinical Neuroscience, 252*(5), 201-209. doi:10.1007/s00406-002-0381-6

Sareen, J., Afifi, T. O., Mcmillan, K. A., & Asmundson, G. J. (2011). Relationship Between Household Income and Mental Disorders. *Archives of General Psychiatry, 68*(4), 419. doi:10.1001/archgenpsychiatry.2011.15

Assari, S., & Caldwell, C. H. (2017). High Risk of Depression in High-Income African American Boys. *Journal of Racial and Ethnic Health Disparities, 5*(4), 808-819. doi:10.1007/s40615-017-0426-1

Mirowsky, J., & Ross, C. E. (1992). Age and Depression. *Journal of Health and Social Behavior, 33*(3), 187. doi:10.2307/2137349

https://www.healthline.com/health-news/people-more-stressed-today-than-1990s#The-bottom-line

Ranking the States. (n.d.). Retrieved December 11, 2020, from https://www.mhanational.org/issues/ranking-states