CHURN REDUNCTION

By C Sai Amogh

Contents:-

1. INTRODUCTION

1.1 Business Problem:-

Churn (loss of customers to competition) is a problem for companies because it is more expensive to acquire a new customer than to keep your existing one from leaving. This problem statement is targeted at enabling churn reduction using analytics concepts. The objective is to predict customer behavior

1.2 Metrics for the Business Problem

1.2.1 Area under the curve(ROC curve) :-

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm.

1.2.2 Confusion matrix :- A confusion matrix is a table that is often used to describe the performance of a classification model and has two dimensions actual and predicted.

1.2.3 Accuracy score :- Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

1.2.4 Classification report :- Precision , recall (Sensitivity) , F1 score

1.3 Data

- account length
- international plan
- voicemail plan
- number of voicemail messages
- total day minutes used
- day calls made

- total day charge
- total evening minutes
- total evening calls
- total evening charge
- total night minutes
- total night calls
- total night charge
- total international minutes used
- total international calls made
- total international charge
- number of customer service calls made

Target Variable :

Move: if the customer has moved (1=yes; 0 = no)

Aim :- To Classify Move

1.4 Sampling technique

1.4.1 Random sampling:- A simple random sample is a subset of a statistical population in which each member of the subset has an equal probability of being chosen here Training and testing is done by random sampling in feature engineering

1.5 :- Mongo DB

Training set is been stored in MongoDB

2 METHODOLOGY

2.1 EDA and Statistical analysis

2.1.1 Train and Test set Dimensions

Train set:-

There are 3333 rows and 22 variables

Test set:

There are 1667 rows and 22 variables

2.1.2 Missing values and Categorical data distribution:-

The data has no Missing values

Categorical variables in the dataset are

1. Churn
2. State
3. Phone Number
4. International plan
5. voice mail plan

From the above we can see that there are 5 Categorical variables

- Unique Points in categorical variables are as followers

  - unique points in phone number: 3333
  - unique points in international plan: 2
  - unique points in state: 51
  - unique points in voice mail plan: 2

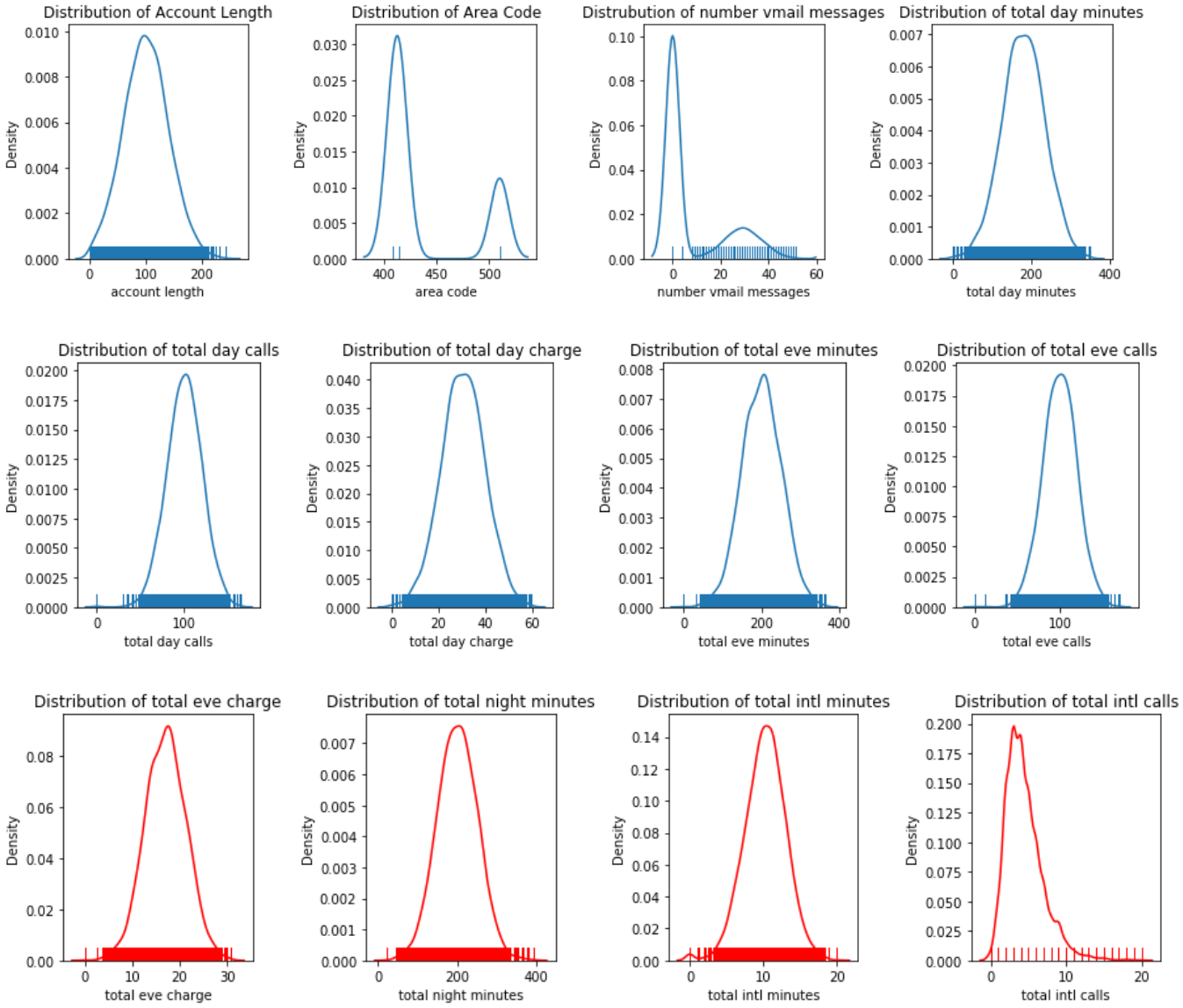
- Distribution of points in churn (Ture or False)

  False. 2850
  True. 483

- From the above we can see that dependent variable is imbalanced.

2.1.3 Density Plots on independent variables

Density plots are great way to visualize the distribution of the numerical data and they are alternative to histograms further validate the statistical analysis

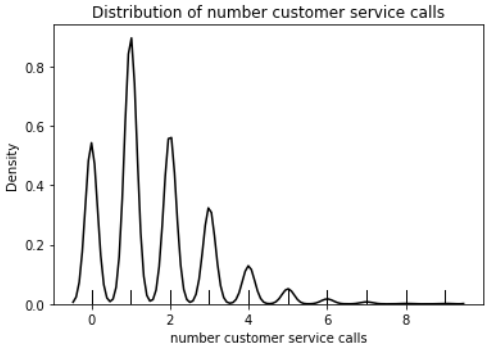


Findings from the above density plots

- We can see that most of the variables are almost normally distributed

- Variables account length and total day minutes , total evening calls , total night calls are normally distributed this is found by statistical analysis where I found that mean and me dian are the same and this is confirmed by density plots
- Area code has 3 sub variables
- Total international calls is right skewed and this variable might be important

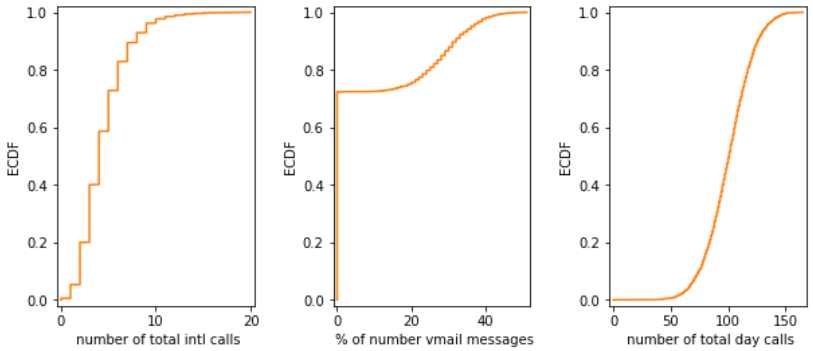
Distribution of customer service calls

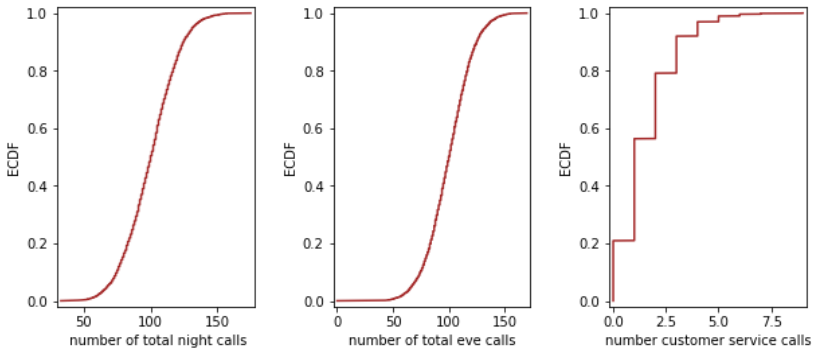


- The above variable has many different sub variable distributions so this variable might be the most important one for prediction

2.1.4 CDF

The cumulative distribution function (cdf) is the probability that the variable takes a value less than or equal to x.

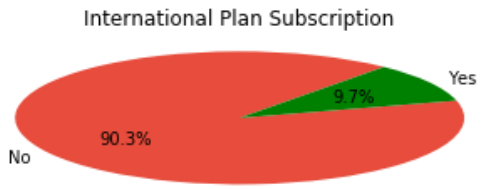




Findings from the above

- If we see the Cumulative distribution we see that about 0.8 international calls are below 10 ,
- From total day calls 60 % of the people are making 100 calls a day
- Total night calls 40 % people are making 100 calls a day at night
- Total eve calls 20 % people are making 50 to 100 calls at evening time
- Service calls 60 to 80 % are making < 3 calls to service center

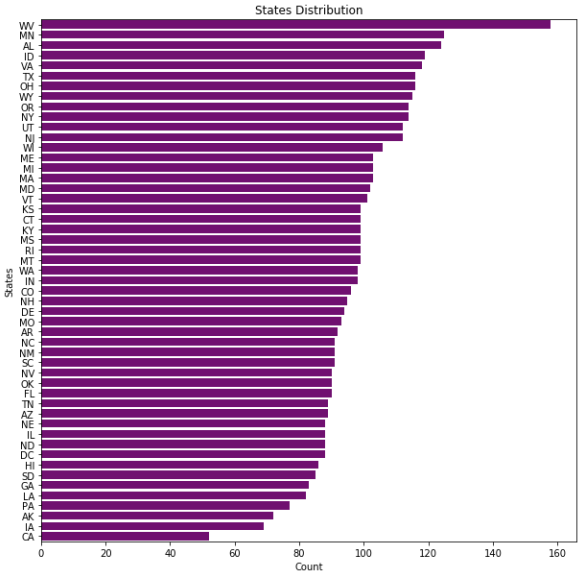
2.1.5 Pie plots for international plan and voice mail plan



From the above we can see that

- In International plan subscription about 9.7 % have taken the plan and rest have not taken the plan
- In Voice mail plan about 27.7 have voice mail plan and rest don't

2.1.6 Bar plot for states



From the bar plot for the state we state WV tops and followed by MN and least is CA

2.1.7 Chi square test for categorical variables

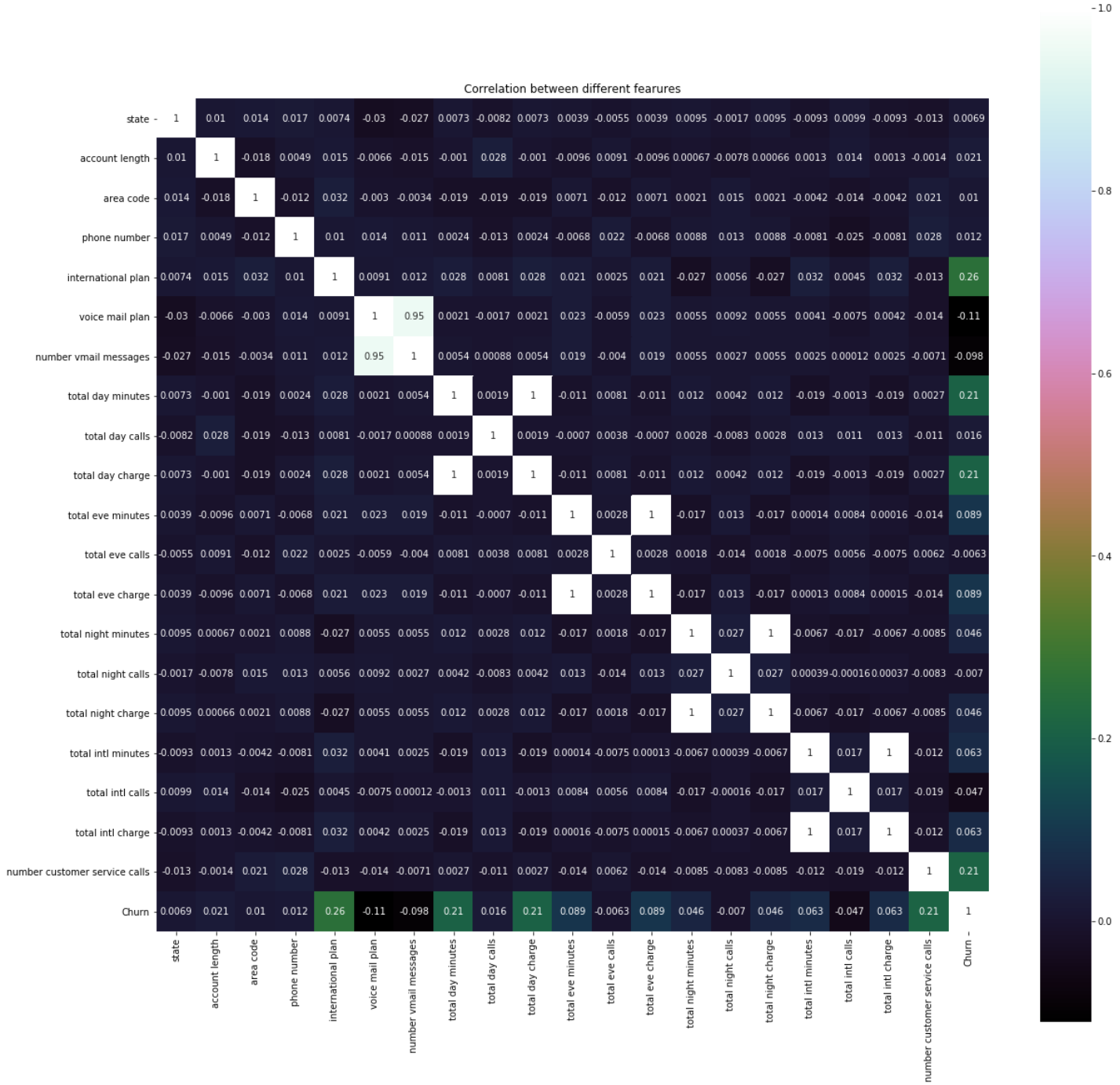
Here I am taking international plan , state, voice mail plan

Results

international plan
2.4931077033159556e-50
state
0.002296221552011188
voice mail plan
5.150639659038 98e-09

- All the three are useful according to chi square test but this will become clear after converting to numerical values and seeing the correlation plot
- Distribution of phone code is unique this can be removed so this is not been included in chi square

2.1.8 Correlation plot



From the above we find the following

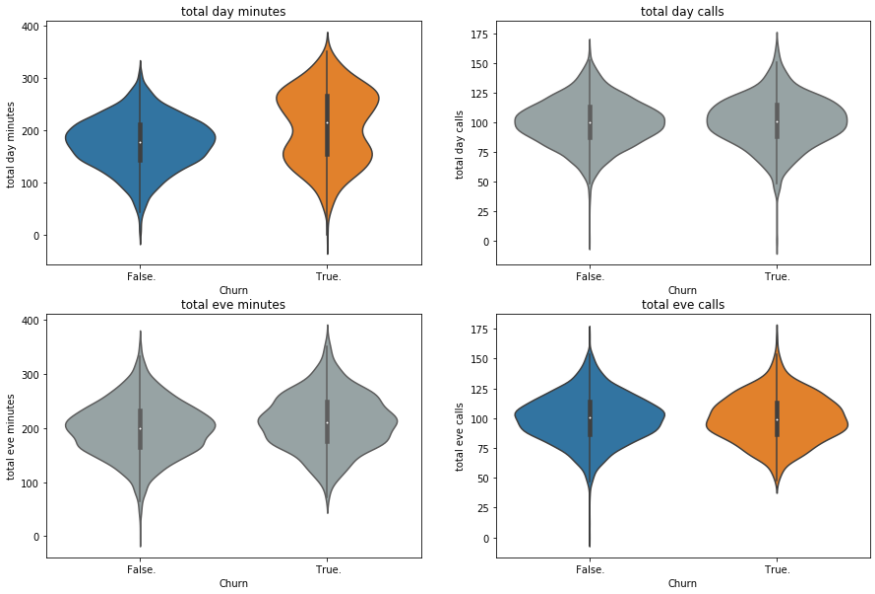
- There is correlation between the following
  - Number of vmails messages and Voice mail plain
  - Total day charge and total day minutes
  - Total eve charge and total eve minutes
  - Total night charge and total night minutes
  - Total int charge and total int minutes

- As there is correlation between there variables one variable will be removed from the variables that are correlated
- I will be removing charge and create new variable called total charge in feature eng

Further findings

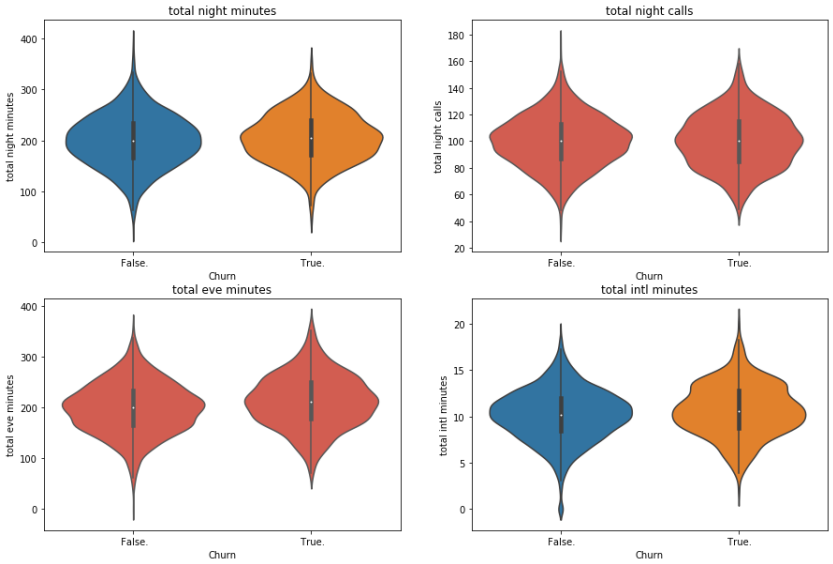
- Correlation between international plan and churn is 0.26 so this an important variable
- Correlation between States and Churn is 0.0069 which is very low
- There is an negative correlation with voice mail and number of vmails messages with churn
- There is also negative correlation b/w eve calls, night calls, int calls with churn
- State can removed as this is not so important
- Further while doing regression there negative correlation variables with churn can removed

2.1.9 Bivariate analysis(Violin plots)

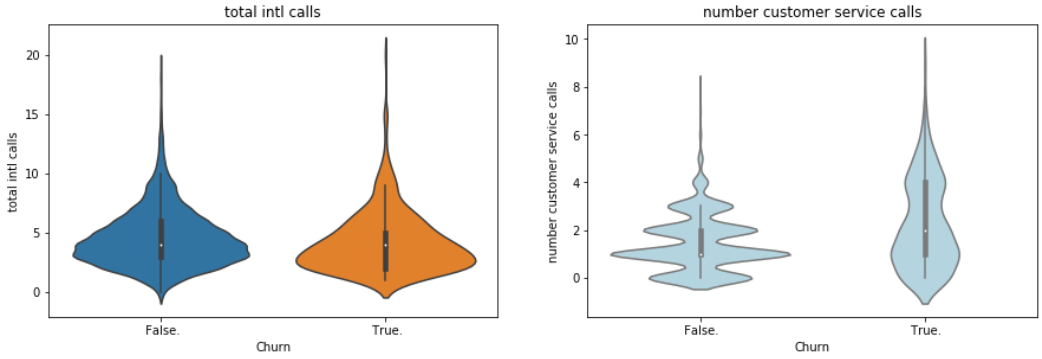


From the above we can see that

- total day minutes :- False has a less width distribution then the True and the length is more then True further talk time is less at the range of 150 to 200 for True and little more for who don't churn out and box plot is larger for true(who churn out)
- total day calls:- box plot is almost same but false has a wider distribution
- total eve minutes:- people who churn out are actually talk in the evening



- we can clearly see the effect on churn in total night minutes , eve minutes



- we can clearly see the effect of customer service calls on churn(True) box plot is higher and distribution is constant and variation in spread, length is seen for churn(True)

2.2 Feature Engineering

2.2.1 Feature variables

- After thorough investigation of variables in data set I have come up with the following feature variables.
- For convince I have combined train and test rows together as train

I would like to give the reasons for choosing the feature variables first they are as follows.

- Total (charge of all the calls) :- As there is correlation b/w charge and minutes , I have created a new variable called total (which is addition of charge of day calls, night, eve, int calls) and removed individual charge

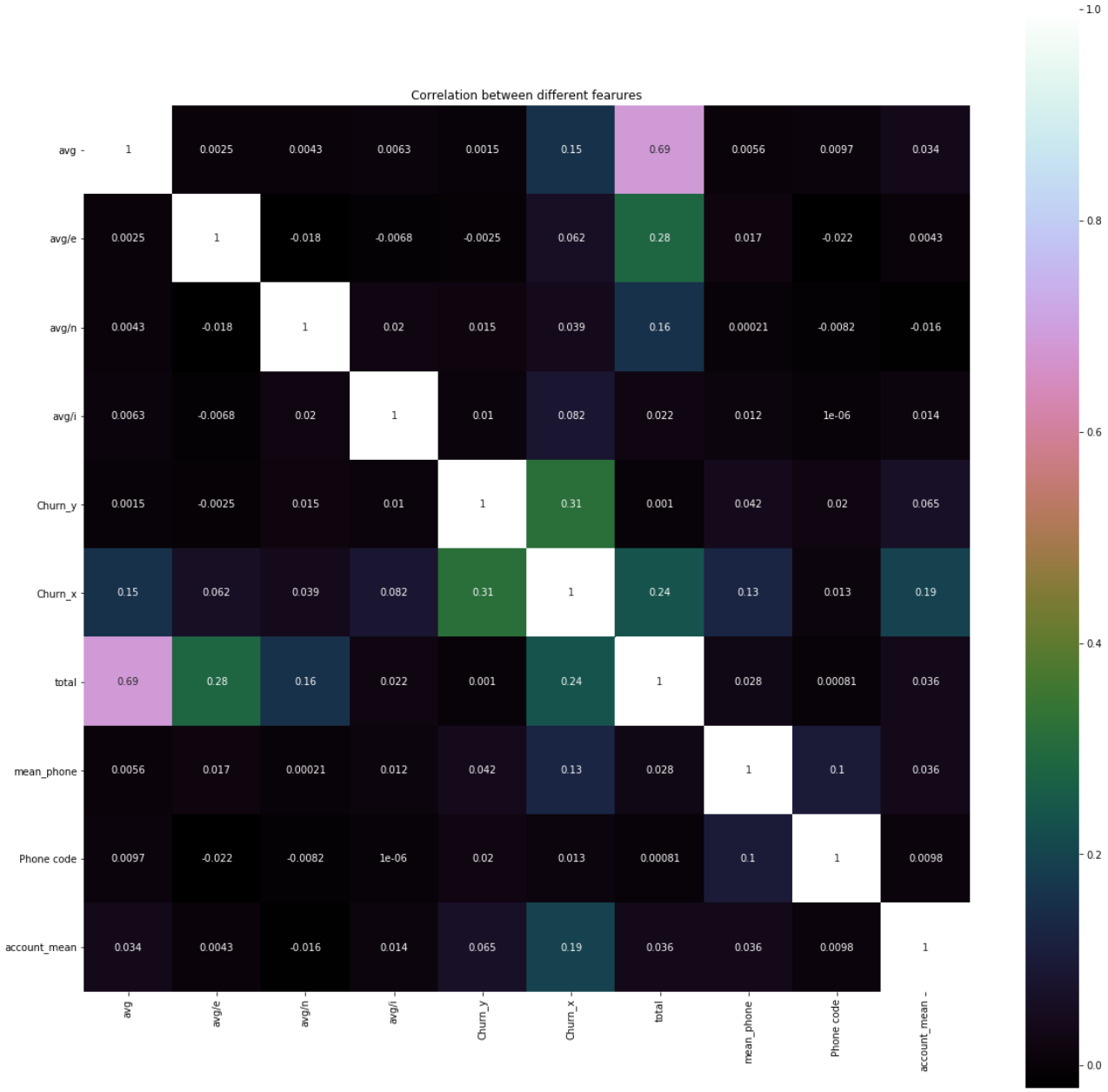
- Phone code mean:- Phone code is an unique categorical variable but has code as common so I have split the number and taken the code and did a group by the phone code and mean by churn and merge on train
- Customer service calls mean:- group by customer service calls and mean by churn
- Account length mean :- there are more than 200 account lengths, I have taken mean according to Churn , group by accounts and merged on train
- Avg :- total day minutes / total day calls
- Avg / e :- total eve minutes / total avg calls
- Avg/ I :- total int minutes / total int calls
- Avg / n :- total night minutes/ total night calls

Following are the feature variables:-

- Total (charge)
- Phone code mean
- Customer service calls mean
- Account length mean
- Avg
- Avg / e
- Avg/ I
- Avg / n

2.2.2 Validation of Features:-

2.2.1 Correlation plot

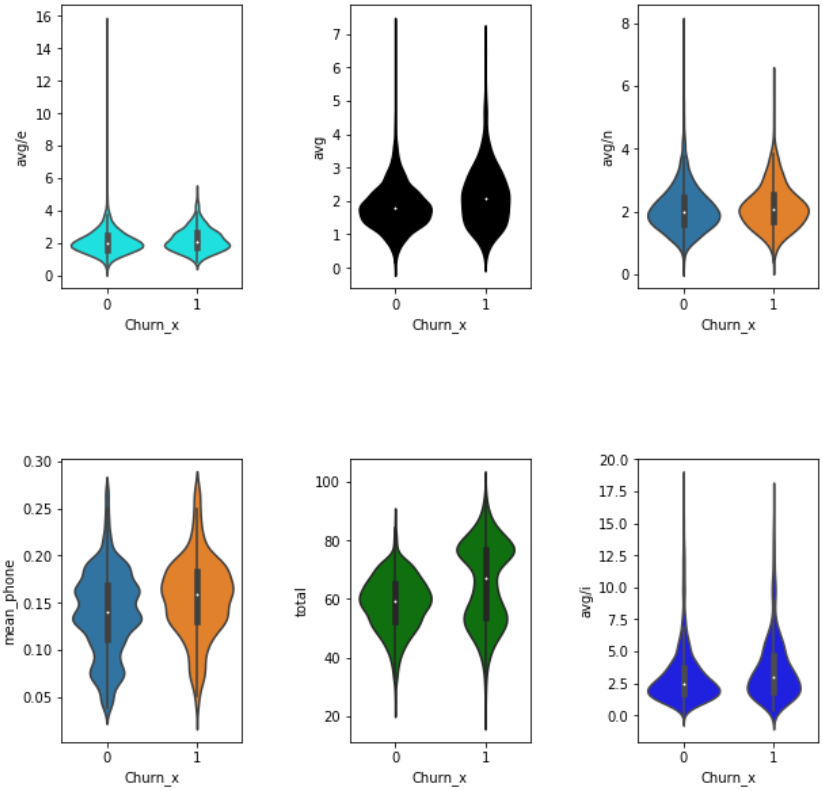


From the correlation plot above we get the following

- Avg/I, avg/n, avg/e might not be very important to predict churn (churn_x) as the correlation is very low
- Total (total of all charge) vs Churn(churn_x) is 0.24 this is one of the =important variable
- Customer service calls (churn_y) vs churn(churn_x) 0.31 good correlation

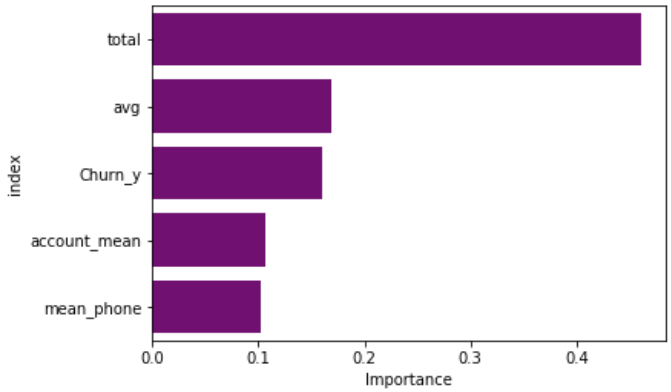
- Account length(account mean),avg,phone code mean(phone_mean) vs churn is 0.19 .0.15,0.13 these 3 variables are effecting churn

2.2.2 Bivariate analysis(violin plots)



- We can clearly see how each feature is affecting churn
- Avg/n for both 0 and 1 distribution is similar and corr is 0.39
- Avg/I for both 0 and 1 box plot for o is higher and distribution is almost similar and corr is 0.69

2.2.3 Random forest feature importance:-

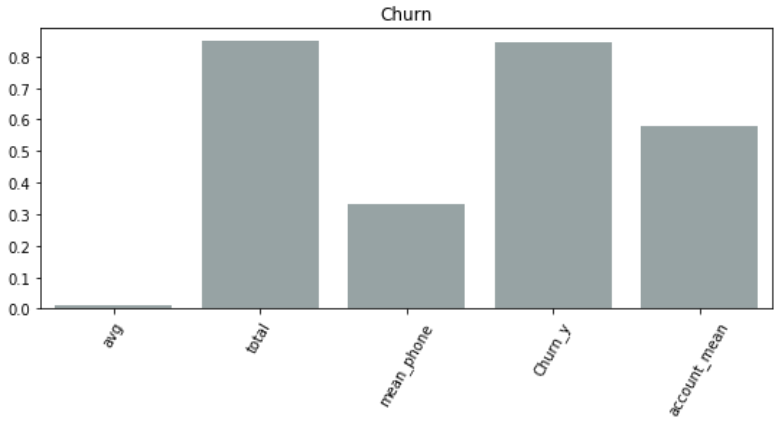


From random forest- total, avg followed by customer service center (churn_y) are contribution about to about 80 % to the model to predict churn

The more the variable appears(mode) the importance of the variable increases in random forest

2.2.4 Feature importance by logistic regression

Weights or coef of the variables give the feature importance, larger the weight, the more importance is the variable.



Customer service call mean (Churn_y) and total are the most important variable.

3 Modeling

This is a classification problem so classification related models will be used

Data Preprocessing :- Dummy variables for area code (city,city_1)

3.1 Logistic Regression:- Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome.

- Data normalization :- Standard Scaler method has been applied for data normalization
- Hyper turning parameter is alpha which is C = 1/alpha , Penalty – l1 or l1 norm (regularization)
- These hyper parameters controls the over fitting and under fitting.

As the data set is imbalance , one method to balance the data is class weights

- class_weight : dict or 'balanced', default: None'
- The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data
- I choose balanced

With the help of grid search and cross validation of 5 following was the best C and penalty

Best Penalty: l1
Best C : 0.8

3.2 SVM: support vector machines (SVM) are supervised learning models with associated learni ng algorithms that analyze data used for classification and regression analysis

Linea SVM was used

- Data normalization :- Standard Scaler method has been applied for data normalization
- Hyper turning parameter is alpha which is C = 1/alpha
- These hyper parameters controls the over fitting and under fitting.

Best C from CV was {'C': 12}, class weight:- balanced

3.3 Random forest:- Random forests or random decision forests are an ensemble learning meth od for classification, regression and other tasks, that operate by constructing a multitude of decisi on trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

- Here all almost all variables are used as random forest or tree based models are not effect ed by correlation

- Imbalance is taken care by class weight taking as balanced

Hyper tuning of parameters is done by grid search the following are the results

{'n_estimators': 400} (number of decision trees)
{'max_depth': 8} (depth of the tree)
{'min_samples_split': 15}
{'min_samples_leaf': 1}
{'max_leaf_nodes': None}
{'max_features': 0.5}

3.4 XGboost :- XGBoost is an implementation of gradient boosted decision trees designed for s peed and performance.

- Imbalance is taken care by class weight taking as balanced

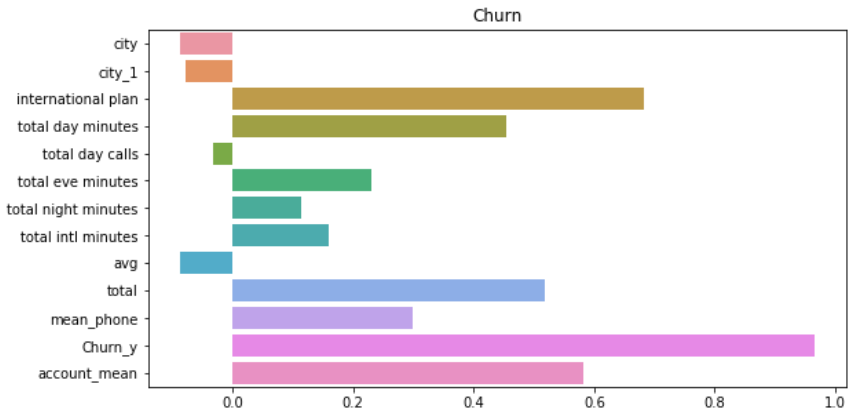
Hyper tuning of parameters is done by grid search the following are the results

{'n_estimators': 100}
{'max_depth': 3}
{'max_depth': 3, 'min_child_weight': 3}
{'gamma': 0.0}
{'subsample': 0.9}
learning_rate=0.05

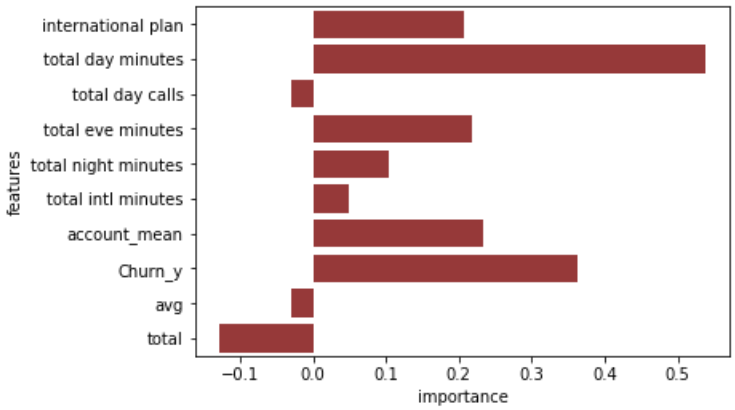
2.4 Feature importance form the models

Interpretability of the model is very important so for this reason, a visualization of features that are important for classification was done and is shown below

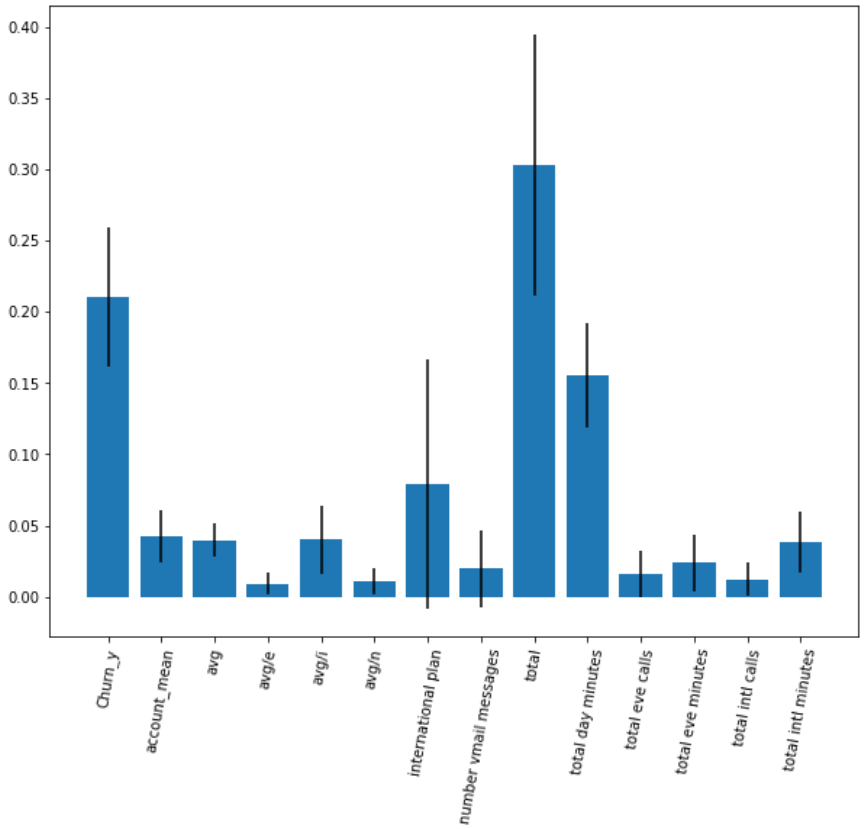
Logistic regression :- Feature importance



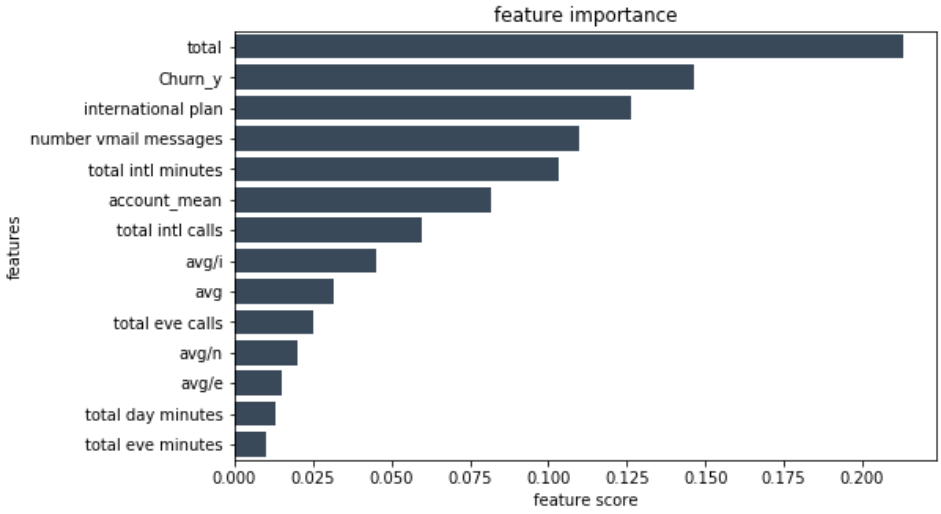
Linear SVM:- Feature Importance



Random Forest:- Feature importance



XGBoost:- Feature importance



3 Results and conclusions

3.1 results

3.1.1 Logistic Regression :-

Roc auc score :- 0.843

Accuracy :- 84 %

Confusion matric :-

[[1209 234]
[34 190]]

Classification Report

	precision	recall	f1-score	support
0	0.97	0.84	0.90	1443
1	0.45	0.85	0.59	224
avg / total	0.90	0.84	0.86	1667

3.1.2 Linear SVM:

Roc auc score:- 0.81

Accuracy :- 82%

Confusion matrix :-

[[1186 257]
[38 186]]

Classification report :-

	precision	recall	f1-score	support
0	0.97	0.82	0.89	1443
1	0.42	0.83	0.56	224
avg / total	0.90	0.82	0.84	1667

3.1.3 Random Forest:-

Roc auc score:- 0.9282249282249282

Accuracy :- 98 %

Confusion matrix :-

```
[[1442       1]
 [  32       192]]
```

Classification report :-

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1443
1	0.99	0.86	0.92	224
avg / total	0.98	0.98	0.98	1667

3.1.4 XGBoost :-

Roc auc score:- 0.9308035714285714

Accuracy :- 98.1 %

Confusion Matrix :-

```
[[1443       0]
 [  31       193]]
```

Classification Report:-

	precision	recall	f1-score	support
0	0.98	1.00	0.99	1443
1	1.00	0.86	0.93	224
avg / total	0.98	0.98	0.98	1667

3.2 Conclusion:- From the results we can see that XGBoost has performed well followed by ran dom forest and feature engineering has helped in classification by reducing the error.