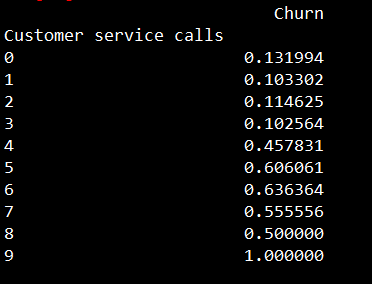
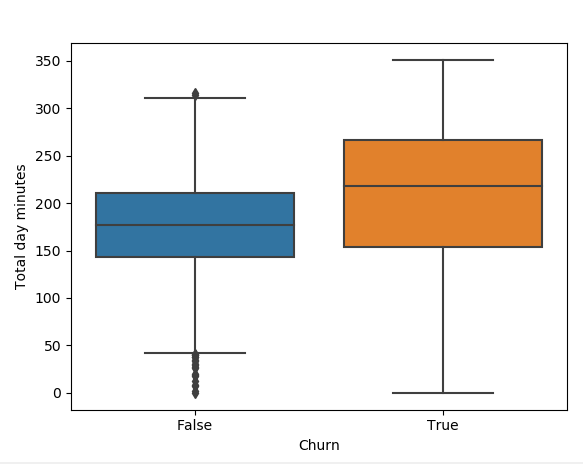
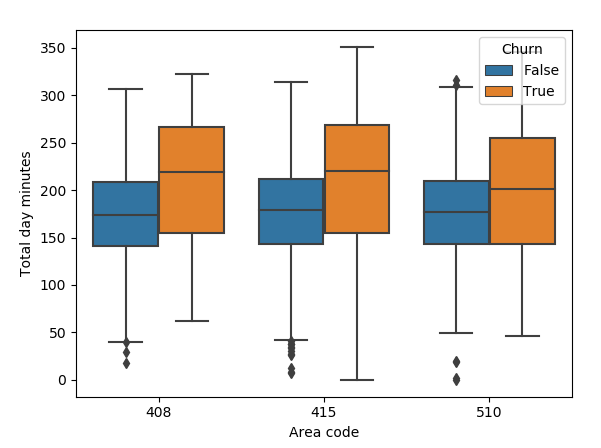
1. Total Rows: 3333
2. No nulls in any column
3. 14% cases are of churn and 86% non churn. That means our model needs to give accuracy better than 86% with good recall
4. Columns:
   1. Customer Service Calls: More the call more the chances of churning



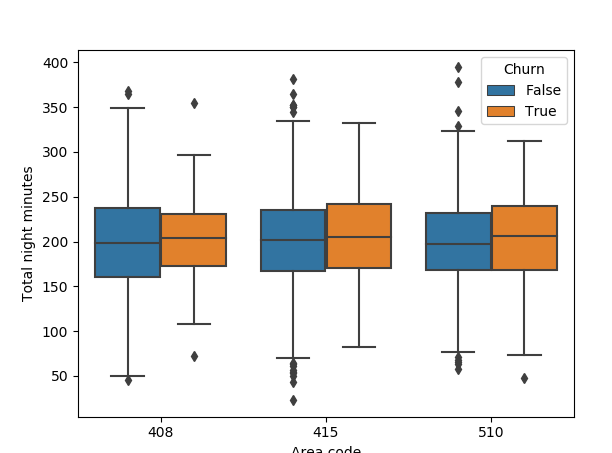
* 1. Area Codes: Total of 3 area codes
  2. Total day minutes vs churn: It looks like “Total day minutes” have huge variance for churned customers, but there could be a lurking variable behind this



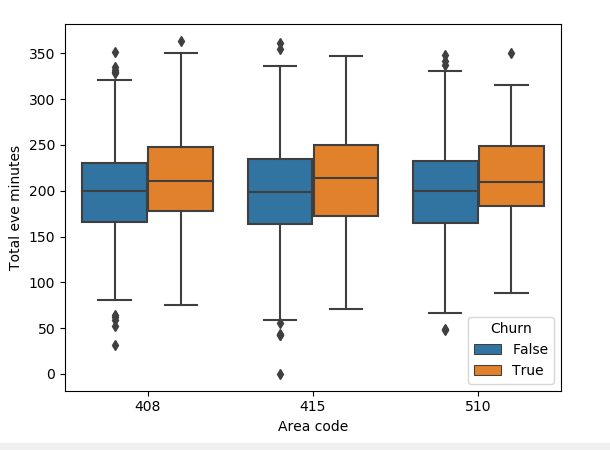
* 1. Area codes vs churn vs Total day minutes: Looks like two area codes have more variance in Total day minutes across churn categories. Need to drill more on this



* 1. Area codes vs churn vs Total night minutes: No major difference

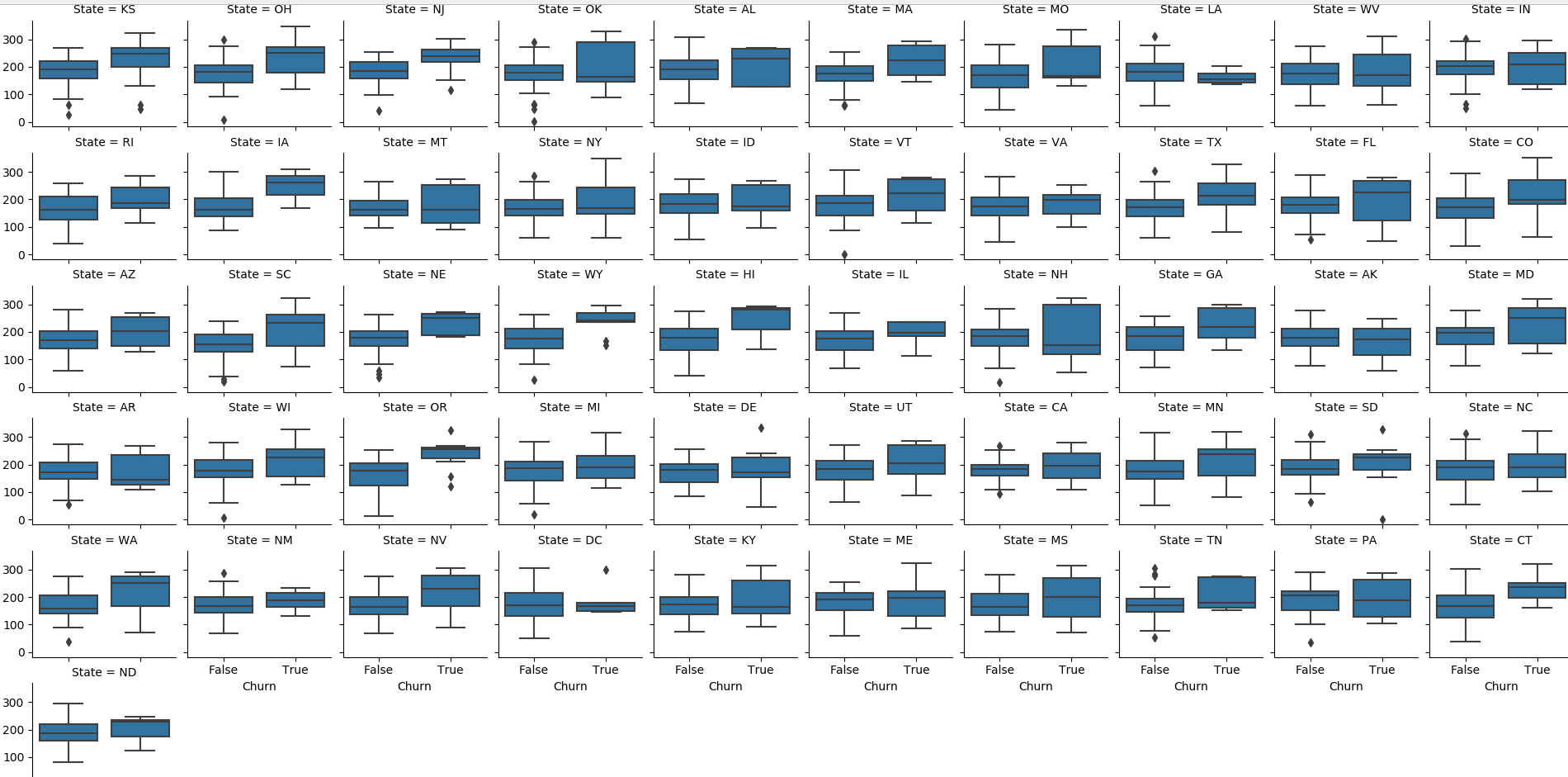


* 1. Area codes vs churn vs Total eve minutes: No major difference

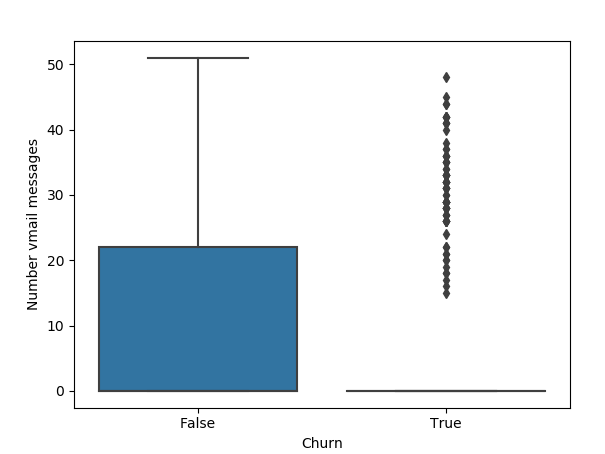


* 1. State: Total 51 states. Out of those 11 states have high ratio for churns

Few states like OR,NJ,IA,WY,HI,OR,WA,DC have a lot of difference in “Total Day minutes” for churned customers as compare to non-churned customers



* 1. Number vmail messages vs Churn: It looks like people who are more likely to churn use very less vmail messages



Models

1. Decision Tree

Accuracy: 93%

ROC\_AUC: 84%

With following aprams

model = DecisionTreeClassifier(class\_weight = "balanced"

,max\_depth = 6

,min\_samples\_split = 8

,random\_state = 6)

1. XGBoost:

Accuracy: 95%

I tried to improve the recall of the models

model = xgb.XGBClassifier(

objective = 'binary:logistic'

,early\_stopping\_rounds = 5

,booster = 'gbtree'

,learning\_rate = 0.1

,max\_depth = 8

,colsample\_bytree = .9

,n\_estimators = 100

,subsample = 0.8

)

Feature Importance

