Churn Reduction

Eeshan Gupta

20 *October* 2018

Contents

Chapter 1

Introduction

Churn rate in its broadest sense, is a measure of the number of individuals or items moving out of a collective group over a specific period. The term is used in many contexts, but is most widely applied in business with respect to a contractual customer base. For instance, it is an important factor for any business with a subscriber-based service model, including mobile telephone networks and pay TV operators.
Churn rate refers to the proportion of contractual customers or subscribers who leave a supplier during a given time period. It is a possible indicator of customer dissatisfaction, cheaper and/or better offers from the competition, more successful sales and/or marketing by the competition, or reasons having to do with the customer life cycle

1.1 Problem Statement

The objective of this case study is to predict customer behaviour. The main study involves around the given usage pattern and the whether the customer has left the business or not. The solution is presented as a machine learning algorithm which predict the churn score based on usage pattern. The predicting features are listed below followed by glimpses of the dataset provided.

```
State                    3333 non-null category
Account_length           3333 non-null int64
Area_code                3333 non-null category
Phone_number             3333 non-null object
Intl_plan                3333 non-null category
Voicemail_plan           3333 non-null category
Number_vmail_message     3333 non-null int64
Day_mins                 3333 non-null float64
Day_calls                3333 non-null int64
Day_charges              3333 non-null float64
Eve_mins                 3333 non-null float64
Eve_calls                3333 non-null int64
Eve_charges              3333 non-null float64
Night_mins               3333 non-null float64
Night_calls              3333 non-null int64
Night_charges            3333 non-null float64
Intl_mins                3333 non-null float64
Intl_calls               3333 non-null int64
Intl_charges             3333 non-null float64
Cust_serv_calls          3333 non-null int64
```

Figure 1.1: Predicting Variables

	State	Account_length	Area_code	Phone_number	Intl_plan	Voicemail_plan	Number_vmail_message
0	16	128	415	382-4657	0	1	25
1	35	107	415	371-7191	0	1	26
2	31	137	415	358-1921	0	0	0
3	35	84	408	375-9999	1	0	0
4	36	75	415	330-6626	1	0	0

Figure 1.2: Columns 1 to 7

	Day_mins	Day_calls	Day_charges	Eve_mins	Eve_calls	Eve_charges	Night_mins
0	265.1	110	45.07	197.4	99	16.78	244.7
1	161.6	123	27.47	195.5	103	16.62	254.4
2	243.4	114	41.38	121.2	110	10.30	162.6
3	299.4	71	50.90	61.9	88	5.26	196.9
4	166.7	113	28.34	148.3	122	12.61	186.9

Figure 1.3: Columns 8 to 14

	Night_calls	Night_charges	Intl_mins	Intl_calls	Intl_charges	Cust_serv_calls	Churn
0	91	11.01	10.0	3	2.70	1	0
1	103	11.45	13.7	3	3.70	1	0
2	104	7.32	12.2	5	3.29	0	0
3	89	8.86	6.6	7	1.78	2	0
4	121	8.41	10.1	3	2.73	3	0

Figure 1.4: Columns 15 to 21

Chapter 2

Methodology

This section explain the procedures leading to the development of the machine learning model.

2.1 Data

The data was given in two parts, training data and testing data. The *training data* consisted of 3333 instances and the *testing data* consisted of 1667 instances.The usage pattern is divided into 20 features and there is one variable as an indicator of whether the customer has churned or not. The features explain the usage pattern of a customer, who is indicated by a different phone numbers. The feature *states* depict the state of customer, international plan and voice-mail plan is binary feature describing the status of subscription of these services. The rest of the features can be divided into groups of three, containing total calling minutes, total number of calls, and total charges, for day, evening, night, and international(if subscribed). Last feature is the number of customer service calls.

2.2 Pre-processing

The pre-processing in an integral part of the analysis. It is done so that all the anomalies in the data that may reduce the accuracy of the trained model.

2.2.1 Missing Value

There no missing values in the given data.

State	3333
Account_length	3333
Area_code	3333
Phone_number	3333
Intl_plan	3333
Voicemail_plan	3333
Number_vmail_message	3333
Day_mins	3333
Day_calls	3333
Day_charges	3333
Eve_mins	3333
Eve_calls	3333
Eve_charges	3333
Night_mins	3333
Night_calls	3333
Night_charges	3333
Intl_mins	3333
Intl_calls	3333
Intl_charges	3333
Cust_serv_calls	3333
Churn	3333

Figure 2.1: Missing Value

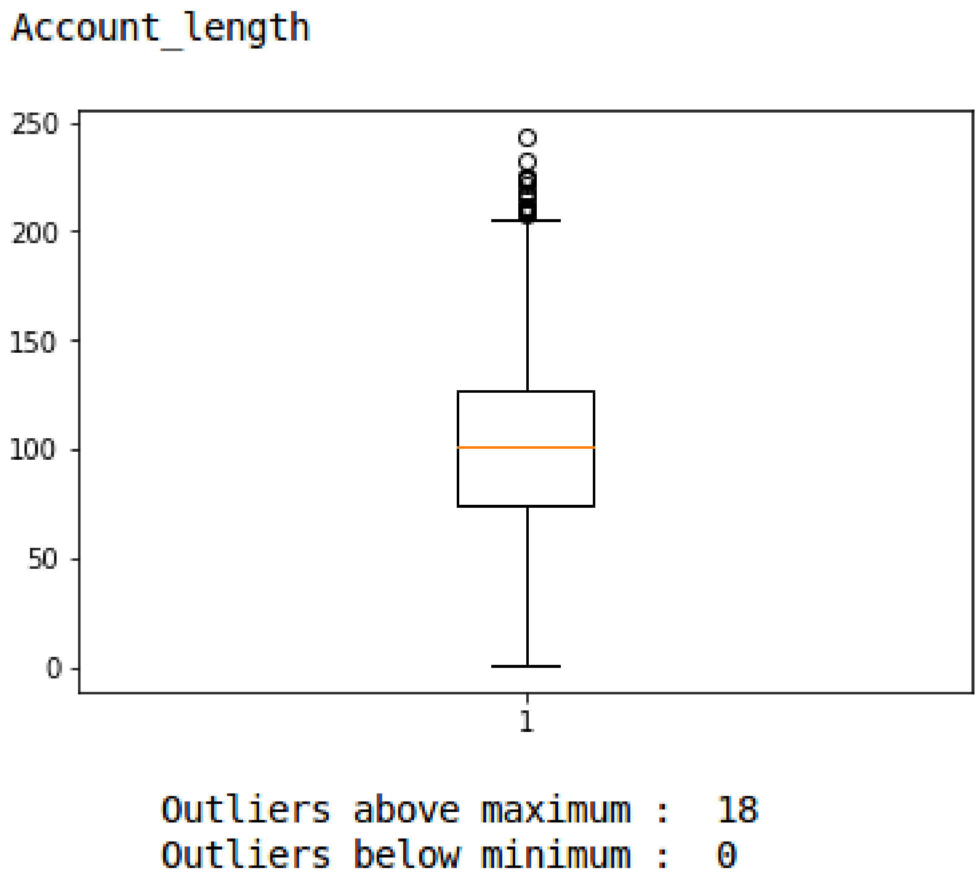
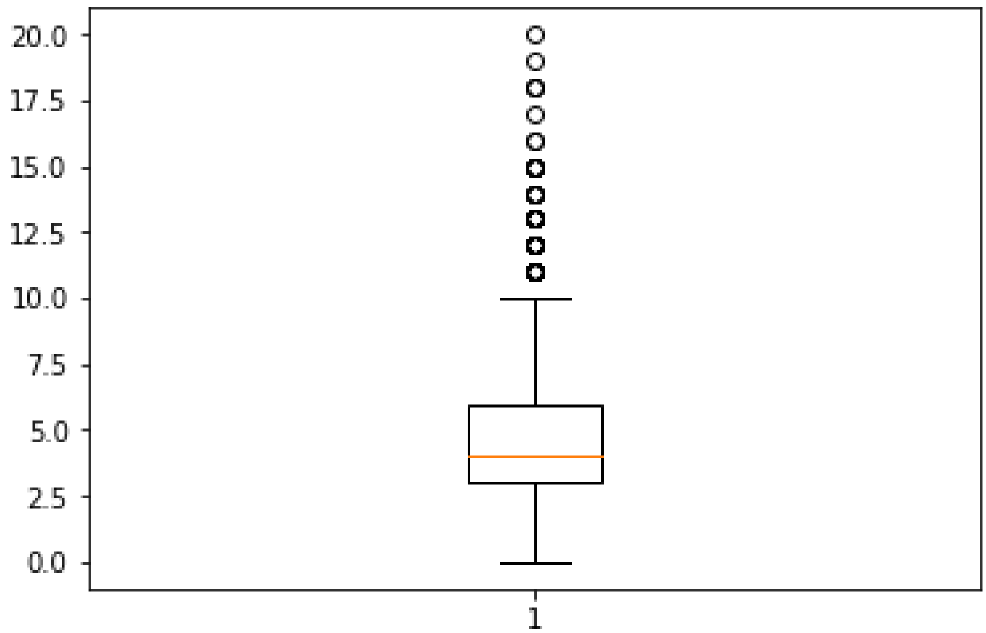


Figure 2.2: Some visualization of outlier analysis

2.2.2 Outliers Analysis

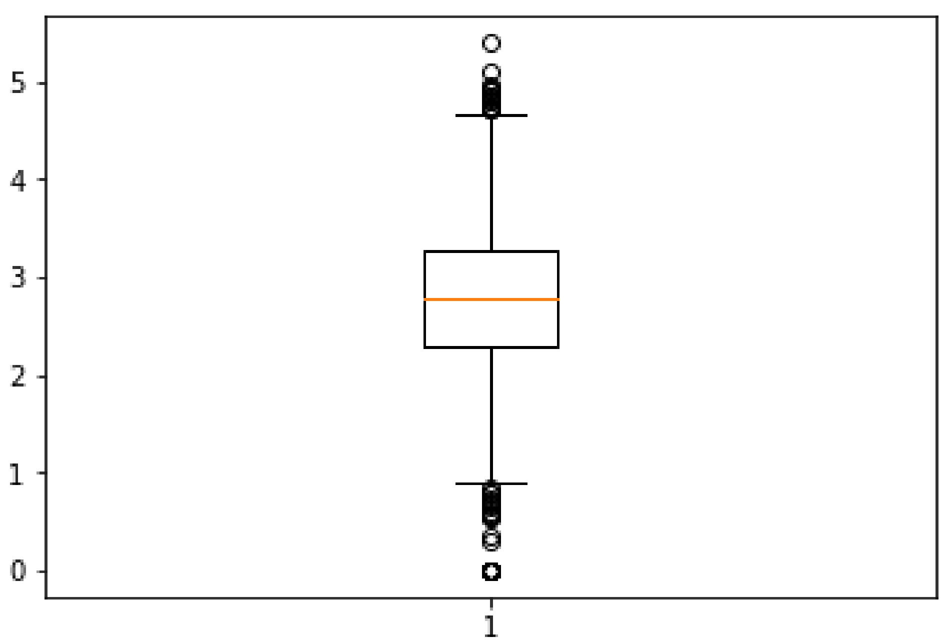
An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses.

Intl_calls




```
Outliers above maximum :  78
Outliers below minimum :  0
```

Intl_charges




```
Outliers above maximum :  17
Outliers below minimum :  32
```

Figure 2.3: Some visualization of outlier analysis

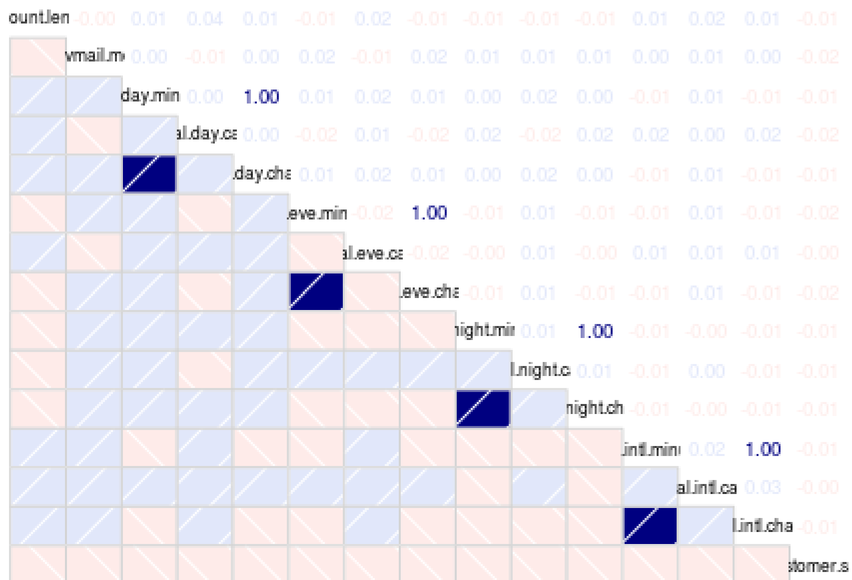


Figure 2.4: Correlation Plot

The outliers are first detected using the box-plot methods. The minimum point, dividing points between 1*st* and 2*nd* quarters and 3*rd* and 4*th* quarter, the median, and the maximum point are calculated. Further the points beyond the maximum point and the points below the minimum point are trimmed to maximum and minimum values, respectively.

2.2.3 Feature Selections

First the features are divided into numerical and categorical types. The correlation plot is drawn between the numerical features of the data.

We can see that there is a high correlation between the total minutes and total charges for day, night, evening, international calls made. The strategy, I employed was to remove total charges features for each type of calls and replace them with charges per minutes for each kind of calls made and created another dataset for it. A copy of original was kept and used for further analysis. The features like area code and phone number are removed from both the datasets as phone number is just used to identify the different customers acting as an ID. The area codes when divided on basis of class, reveals that mean of each numerical value for each class is nearly same and so area code is also not used for the analysis.

Area_code	Churn	Account_length	Number_vmail_message	Day_mins	Day_calls	Day_charges	Eve_mins	Eve_calls	Eve_charges	Night_mins
408	**0**	101.177374	8.025140	172.138408	100.141061	29.263966	199.596159	100.044693	16.965824	198.453980
	1	105.647541	5.573770	206.915574	102.819672	35.176230	211.169262	98.434426	17.949508	203.300410
415	**0**	100.558140	9.038055	176.859302	100.603946	30.066688	198.749471	100.428471	16.893862	201.230162
	1	103.838983	4.262712	210.282415	100.567797	35.748517	212.498093	101.101695	18.062331	206.805508
510	**0**	100.644755	8.323077	175.178182	99.890909	29.780811	199.293566	99.363636	16.940364	199.433077
	1	97.152000	6.280000	199.920400	101.672000	33.987120	212.970800	101.464000	18.102880	204.145200

Day_calls	Day_charges	Eve_mins	Eve_calls	Eve_charges	Night_mins	Night_calls	Night_charges	Intl_mins	Intl_calls	Intl_charges	Cust_serv_calls
100.141061	29.263966	199.596159	100.044693	16.965824	198.453980	99.286313	8.930594	10.099721	4.416201	2.727556	1.386872
102.819672	35.176230	211.169262	98.434426	17.949508	203.300410	97.647541	9.148484	10.423770	4.008197	2.815492	1.942623
100.603946	30.066688	198.749471	100.428471	16.893862	201.230162	100.280479	9.055476	10.274489	4.525018	2.774746	1.419309
100.567797	35.748517	212.498093	101.101695	18.062331	206.805508	101.021186	9.306568	10.866102	4.224576	2.934280	2.042373
99.890909	29.780811	199.293566	99.363636	16.940364	199.433077	100.355245	8.974538	10.073007	4.409790	2.720168	1.492308
101.672000	33.987120	212.970800	101.464000	18.102880	204.145200	101.864000	9.186440	10.634400	3.896000	2.871600	2.088000

Figure 2.5: Table in favour of not choosing Area Code as feature

2.2.4 Feature Scaling

The rest of the numerical features are scaled using *Standardization* technique. The mean of each feature is reduced to 0 and standard deviation is made 1.

2.3 Model Selection

The problem we are dealing with is a classification problem, because the usage pattern of customer will decide whether the he/she will leave the current telecom service provider or not. In the further analysis, there are 6 major classification techniques that are chosen and compared, which are namely:

- Support Vector Classification,
- K-Nearest Neighbours Classification,
- Random Forest Classification,
- Logistic Regression,
- Decision Tree Classification, and
- Naive Bayes Classification

These all algorithm are compared using stratified sampling technique. This sampling technique is chosen specifically as the classes are not distributed equally, rather there is a large difference between the number of users that have not churned and the number of users those who have. The stratified sampling technique is used to remove the inequality of the classes.

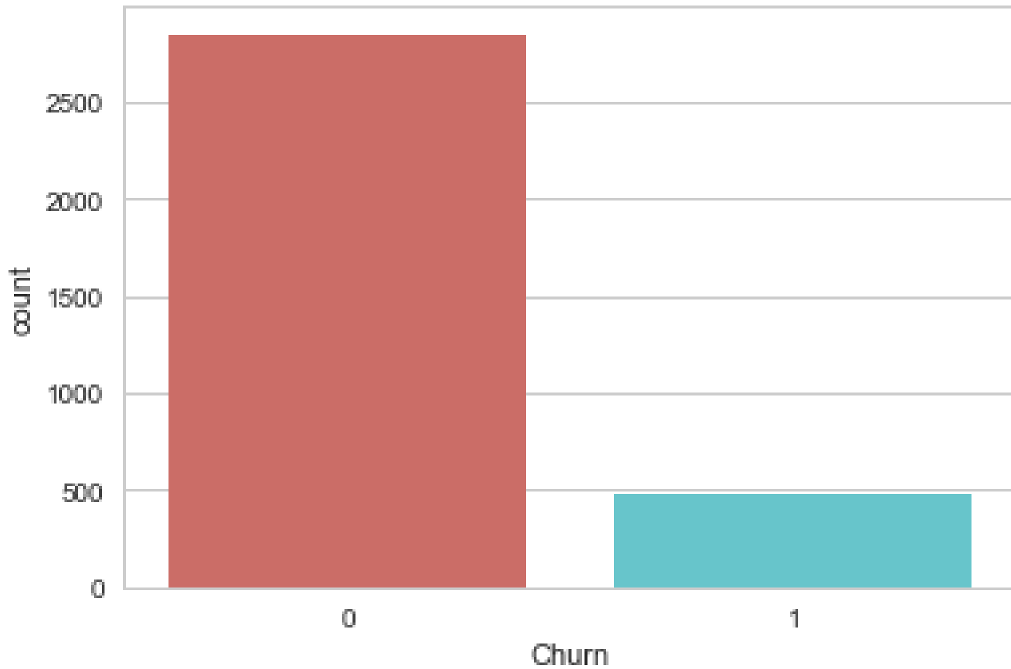


Figure 2.6: Distribution of classes

From the 2.9 we can see the highest accuracy is achieved by *Random Forest Classification* algorithm.
NOTE : The 1st dataset refers to the dataset without charges per minute feature. The 2nd dataset is the dataset is the one with aforementioned feature.
Next the same algorithms are compared for the 2nd dataset. Only results of random forest algorithm are presented here (figure 2.13).

```
SUPPORT VECTOR CLASSIFIER
Accuracy Score  0.916

Confusion Matrix
 [[2821   29]
 [ 252  231]]
```

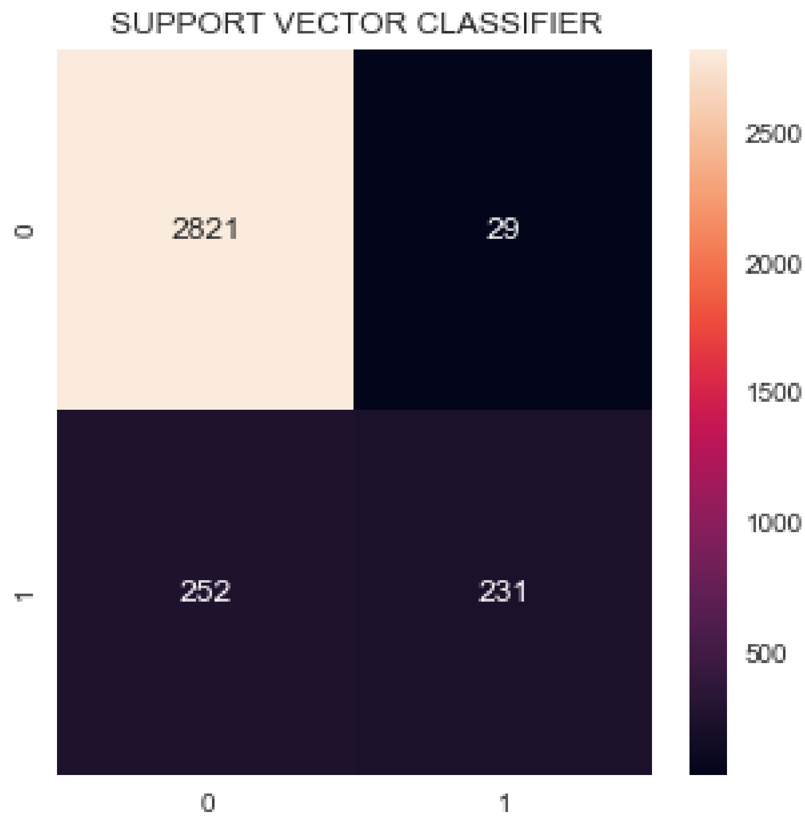


Figure 2.7: Support Vector Classification for 1st dataset

```
K NEAREST NEIGHBOURS CLASSIFICATION
Accuracy Score  0.887

Confusion Matrix
 [[2796   54]
 [ 321  162]]
```

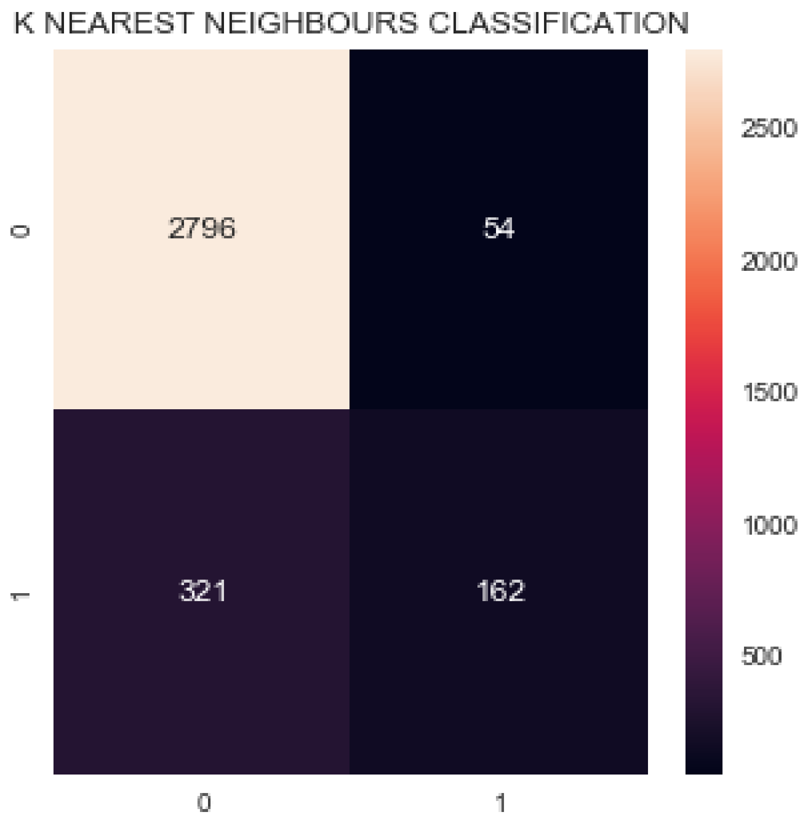


Figure 2.8: KNN Classification for 1st dataset

```
RANDOM FOREST CLASSIFICATION
Accuracy Score  0.942

Confusion Matrix
 [[2821   29]
 [ 165  318]]
```

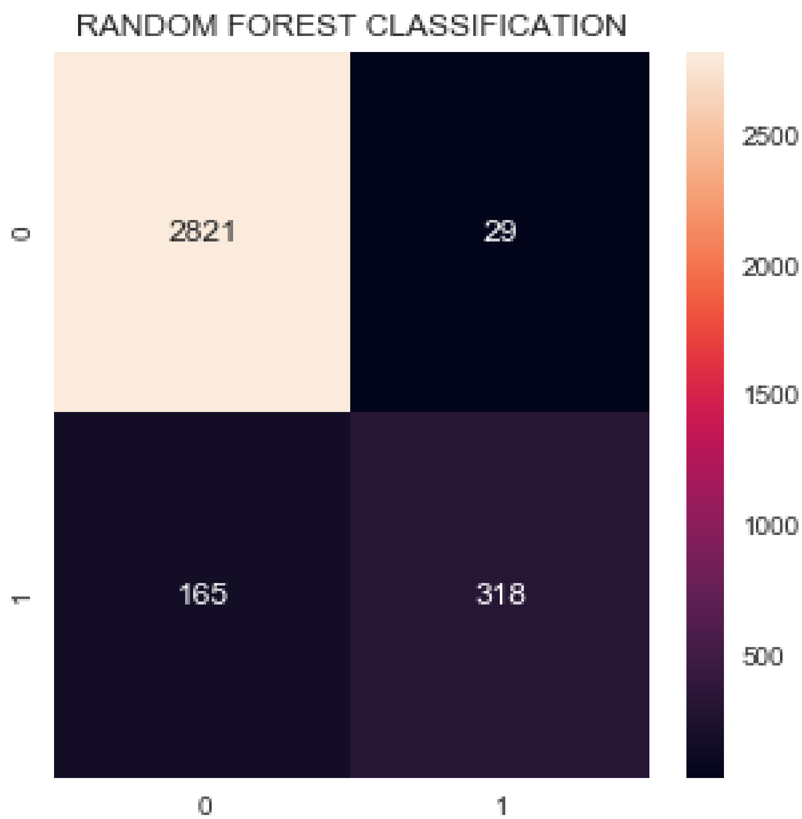


Figure 2.9: Random Forest Classification for 1st dataset

```
LOGISTIC REGRESSION
Accuracy Score  0.863

Confusion Matrix
 [[2778   72]
 [ 383  100]]
```

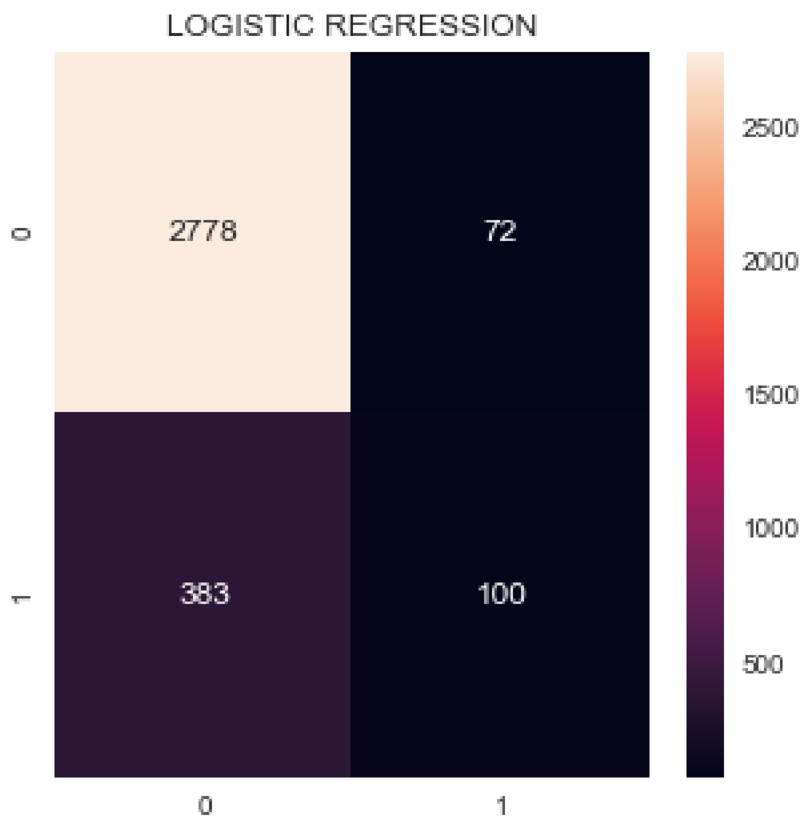


Figure 2.10: Logistic Regression for 1st dataset

```
GAUSSIAN NAIVE BAYES CLASSIFICATION
Accuracy Score  0.866

Confusion Matrix
 [[2645  205]
 [ 241  242]]
```

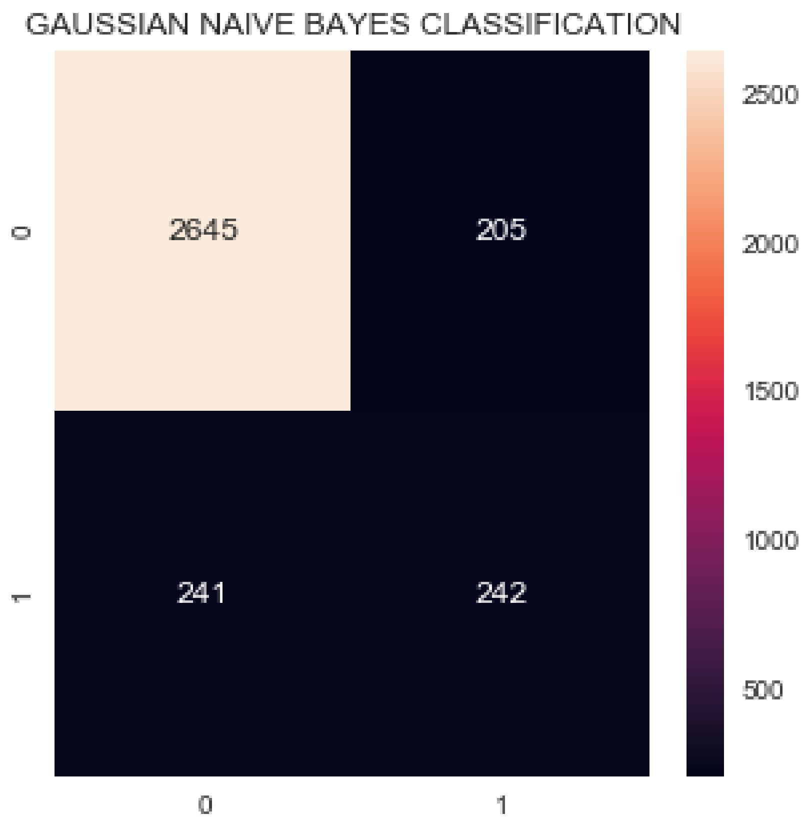


Figure 2.11: Naive Bayes Classification for 1st dataset

```
DECISION TREE CLASSIFICATION
Accuracy Score  0.911

Confusion Matrix
 [[2687  163]
 [ 132  351]]
```

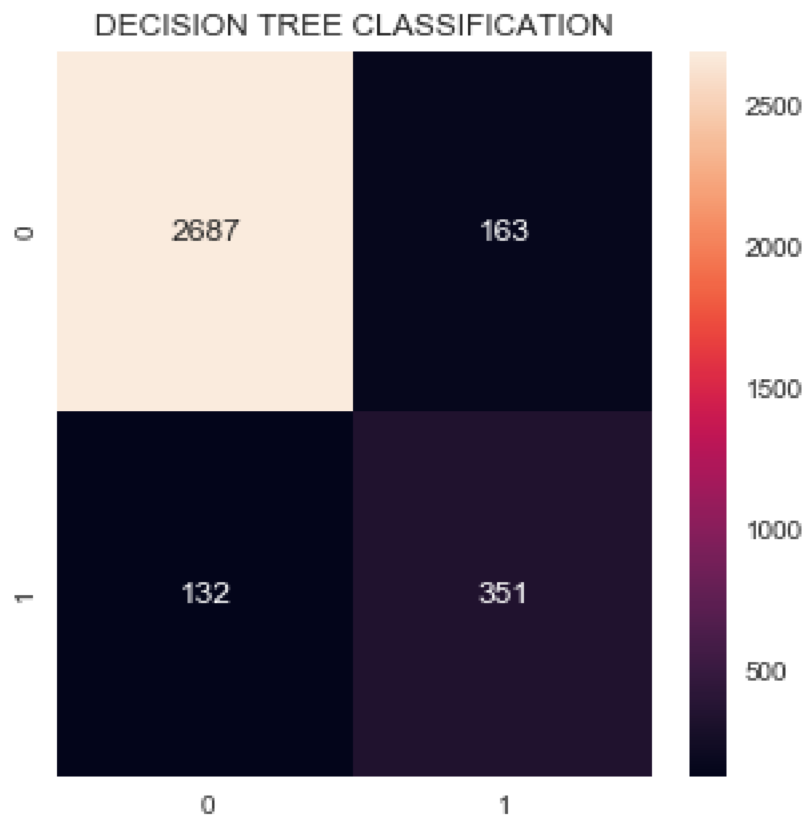


Figure 2.12: Decision Tree Classification for 1st dataset

```
RANDOM FOREST CLASSIFICATION
Accuracy Score  0.935

Confusion Matrix
 [[2820   30]
 [ 185  298]]
```

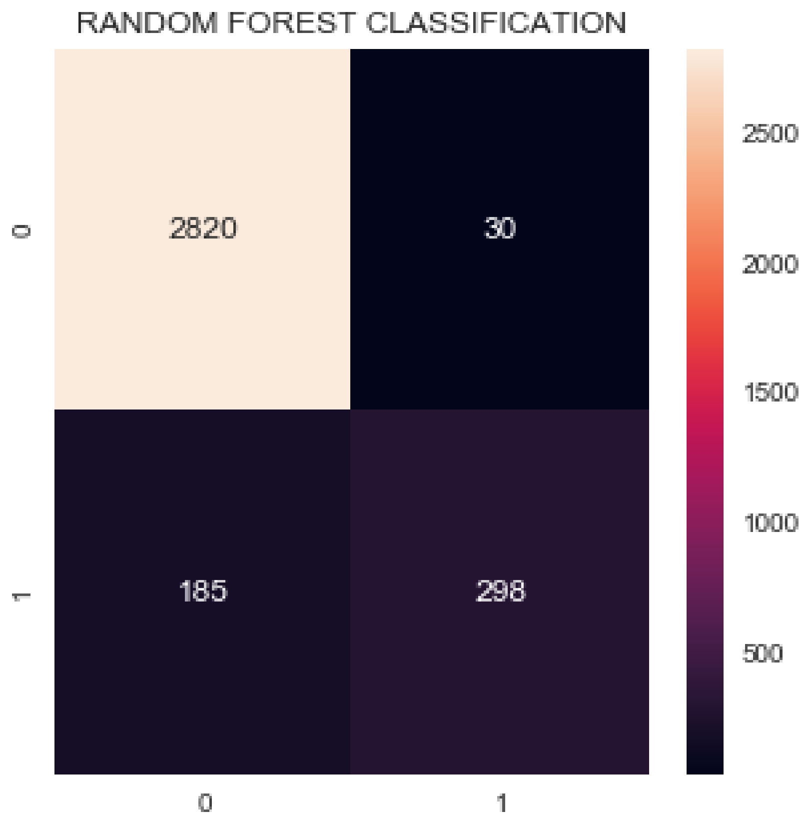


Figure 2.13: Random Forest Classification for 2nd dataset

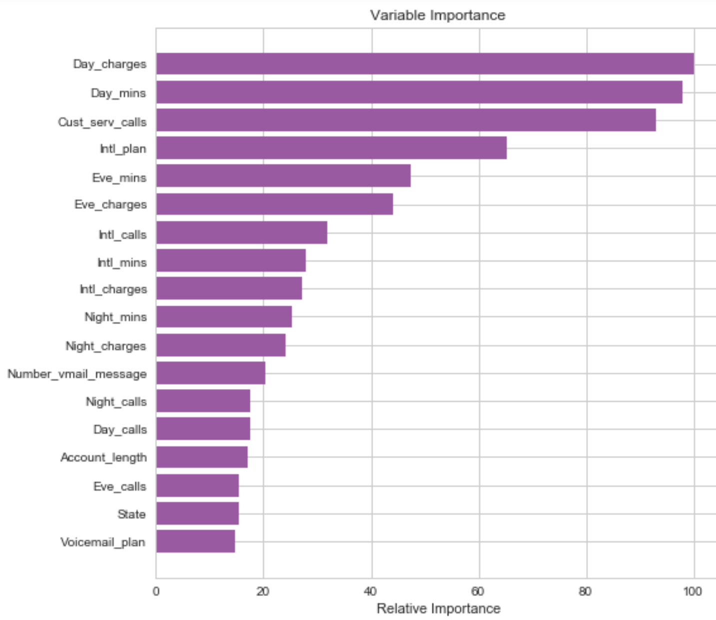


Figure 2.14: Feature dependence

The model is trained with random forest algorithm using 3 different number of decision trees, first with 500 trees, next with 200 trees, and last with 100 trees. The dataset used to train the model is the 1st one as the accuracy of all the algorithms for the 1st dataset is higher than that of the 2nd dataset.

The random forest with 100 trees yielded the best accuracy. The following figure explains the dependence of churn of the features.