BAN 620 Data Mining Project Report

Title: Customer Churn Prediction for a Telecom Company

Submitted by: Group 4

Name	**NETID**
Jasvitha Buggana	js3225
Krupa Shah	yc4954
Manan Upadhyay	rs6739
Preksha Shah	cz2412
Shivani Agrawal	lw3758

1. Objective: Our objective is to develop a predictive model addressing vital aspects of customer churn in phone services data. Employing data visualization techniques, we aim to unveil inherent patterns and trends within the dataset. Simultaneously, we endeavor to identify significant predictors influencing customer churn through various classification techniques. The primary goal is to establish a classification model capable of accurately determining whether a customer is likely to churn or remain, facilitating proactive measures for effective customer retention. Through these comprehensive efforts, we aim to deepen our understanding of customer behavior and recommend strategies that mitigate churn, thereby enhancing overall customer satisfaction.

2. Dataset: https://platform.stratascratch.com/data-projects/customer-churn-prediction

We utilized a data-science practice platform (like leetcode) called Stratascratch to get the dataset, and information about meta data. The dataset contains **3333 rows and 21 columns** and there were no missing values in the data.

```
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   state                   3333 non-null   object
 1   account_length          3333 non-null   int64
 2   area_code               3333 non-null   int64
 3   phone_number            3333 non-null   object
 4   international_plan      3333 non-null   object
 5   voice_mail_plan         3333 non-null   object
 6   number_vmail_messages   3333 non-null   int64
 7   total_day_minutes       3333 non-null   float64
 8   total_day_calls         3333 non-null   int64
 9   total_day_charge        3333 non-null   float64
 10  total_eve_minutes       3333 non-null   float64
 11  total_eve_calls         3333 non-null   int64
 12  total_eve_charge        3333 non-null   float64
 13  total_night_minutes     3333 non-null   float64
 14  total_night_calls       3333 non-null   int64
 15  total_night_charge      3333 non-null   float64
 16  total_intl_minutes      3333 non-null   float64
 17  total_intl_calls        3333 non-null   int64
 18  total_intl_charge       3333 non-null   float64
 19  customer_service_calls  3333 non-null   int64
 20  churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(4)
memory usage: 524.2+ KB
```

3. Description of the dataset: The dataset represented a historical record of a telecommunications business, with a Boolean outcome variable indicating customer churn. A value of 1 for "Churn" signified that the customer had churned, i.e., they had ceased using the company's services. On the other hand, a value of 0 indicated that the customer was still actively using the company's services.

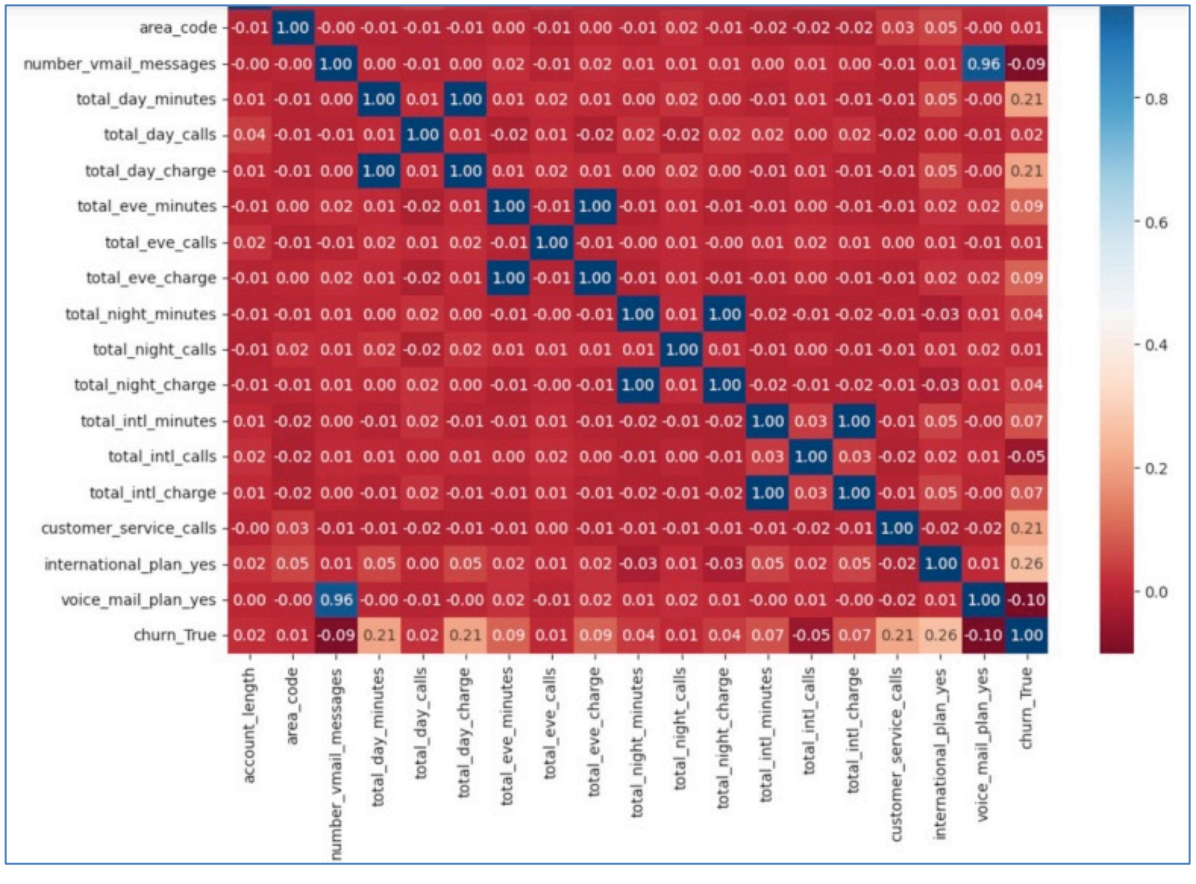
4. Data Preprocessing:

- Changed Outcome Variable: Converted the "Churn" variable from text to a binary format, where "Churn = 1" signifies customers who have churned (stopped using the company's service), and "Churn = 0" denotes customers still using the service.

- Categorized Variables: Identified and categorized variables into numerical and non-numerical types. Numerical variables include integers and floats, while non-numerical variables are represented as strings or objects.

- State Encoding: Coded state names to numerical values for efficient analysis. For instance, "CA" was transformed into the numeric value 4.
- DataType Conversion to Category:Converted 'state', 'phone_number', 'international_plan', 'voice_mail_plan' variables into the category data type for improved memory efficiency and faster processing.
- Encoding Name Change: Renamed variables to enhance clarity and maintain consistency. For example, "Account Length" was changed to "account_length."
- Variable Name Formatting: Removed spaces from the original variable names, ensuring uniformity and simplifying subsequent data handling.

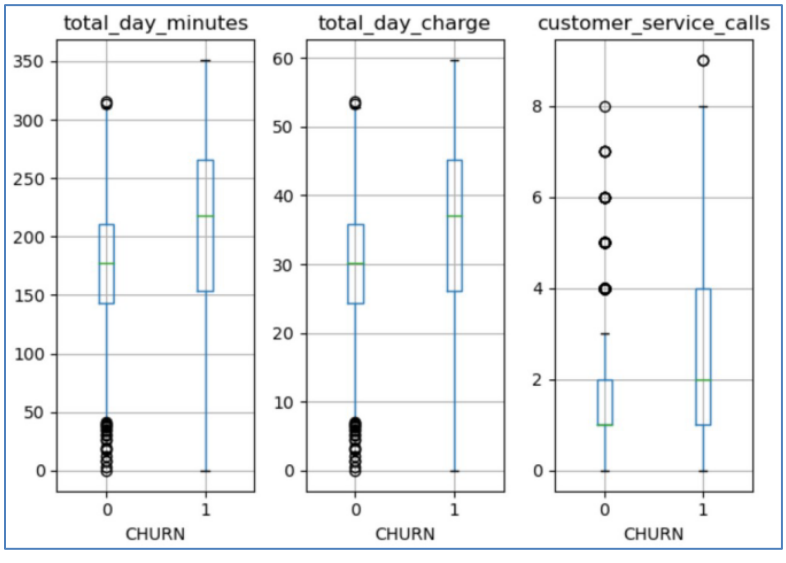
Heatmap:

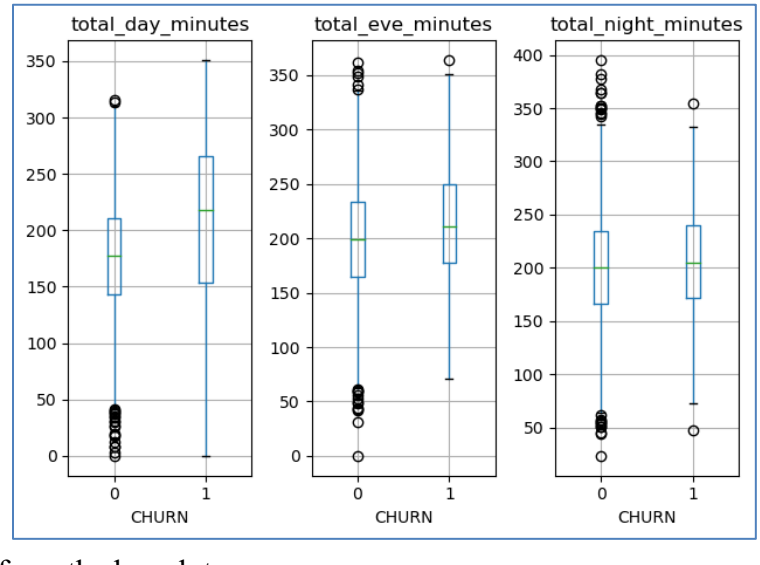


Observing the HeatMap, it was evident that it did not provide strong information regarding variables highly correlating with each other. The strong correlations shown on the heatmap were self-explanatory and couldn't be taken into consideration as they indicated correlation between dependent variables. Consequently, it can be concluded that there was no Multicollinearity present in the dataset.

Since no correlating data was found using the heatmaps, the analysis was initiated on variables like 'total_day_minutes' and 'customer_service_calls,' 'total_day_charge' alongside churn, aiming to extract hints on the churning of customers.

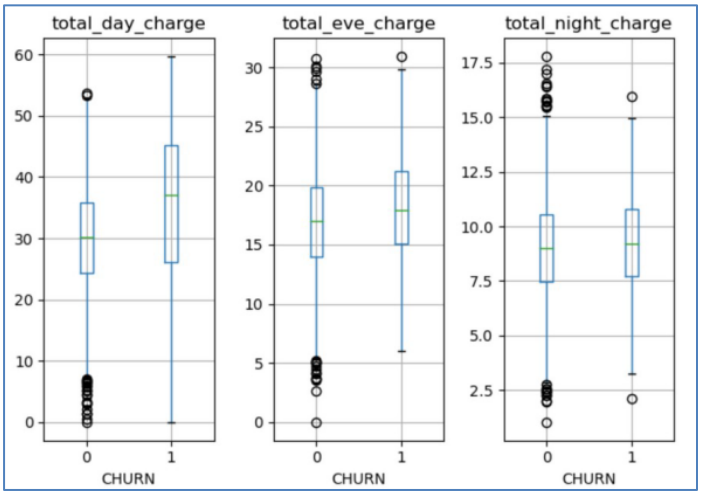
Box Plots:





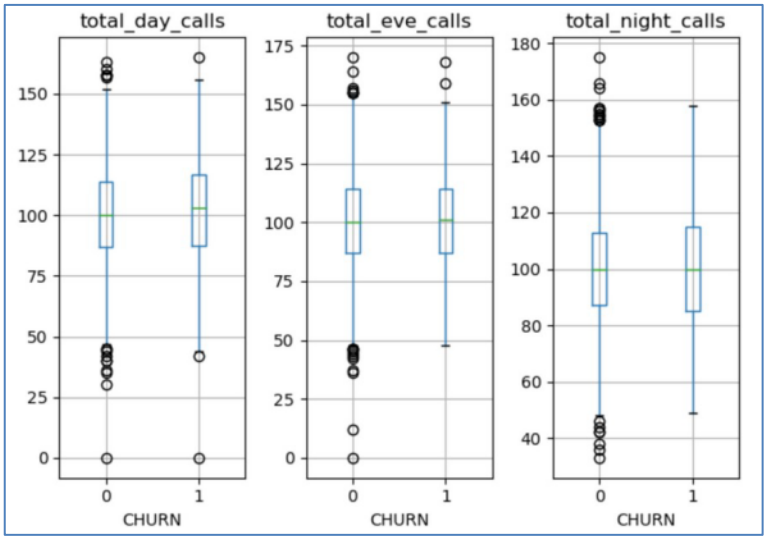
Key insights from the boxplot:

1. Churned customers exhibited a high number of customer service calls, indicating multiple complaints or unresolved issues.
2. Additionally, it is noteworthy that customers who churned tended to pay more than the average due to increased talk time. This observation is significant as it suggests a potential competitive disadvantage, where other telecom companies may offer unlimited plans for a fixed price, while our company charges based on the duration of calls.
3. Daytime experienced the longest call durations for both churned and non-churned customers. Among daytime callers, especially those who churned, conversations were notably lengthier.



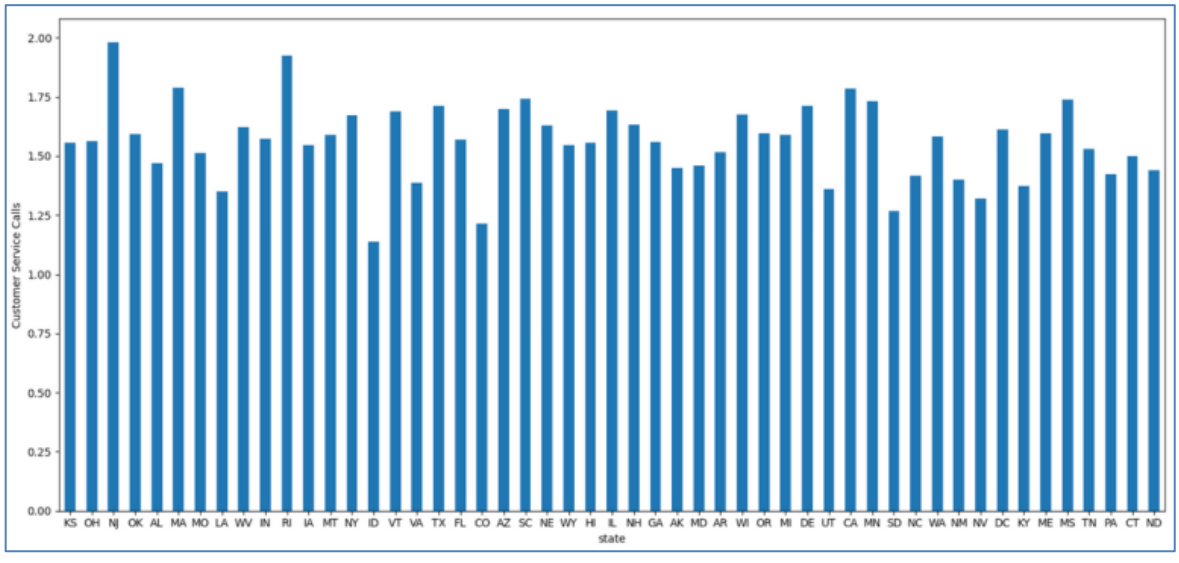
4. An additional observation indicates that churned customers talked more during the day, despite having a similar call count. The key distinction lies in the longer duration of their calls. Furthermore, customers who made daytime calls and incurred charges exceeding $35 were found to be more prone to churn, as observed by this boxplot.

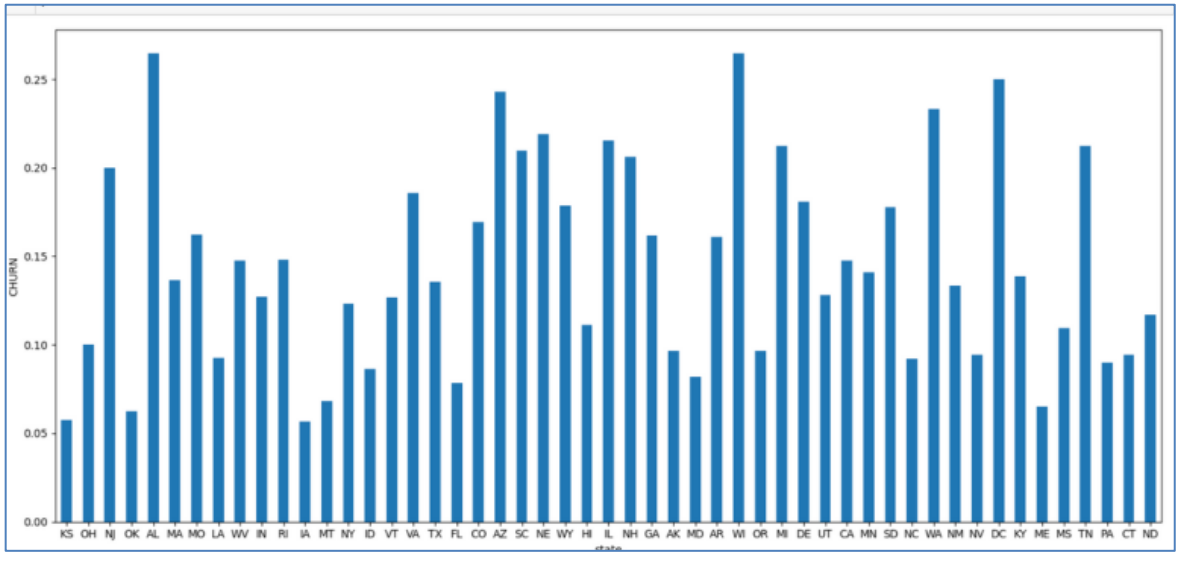
Despite an equal call count, daytime charges were higher, indicating that individuals who churned had longer conversations and made greater payments.



The above box plot reveals insightful information. Upon close examination of the outliers, it becomes evident that customers who engage in more than 150 calls and those who make fewer than 50 calls are less likely to churn. This suggests a higher satisfaction level among these customer segments. The challenge appears to be concentrated among customers who make an average number of calls.

Bar Graphs:

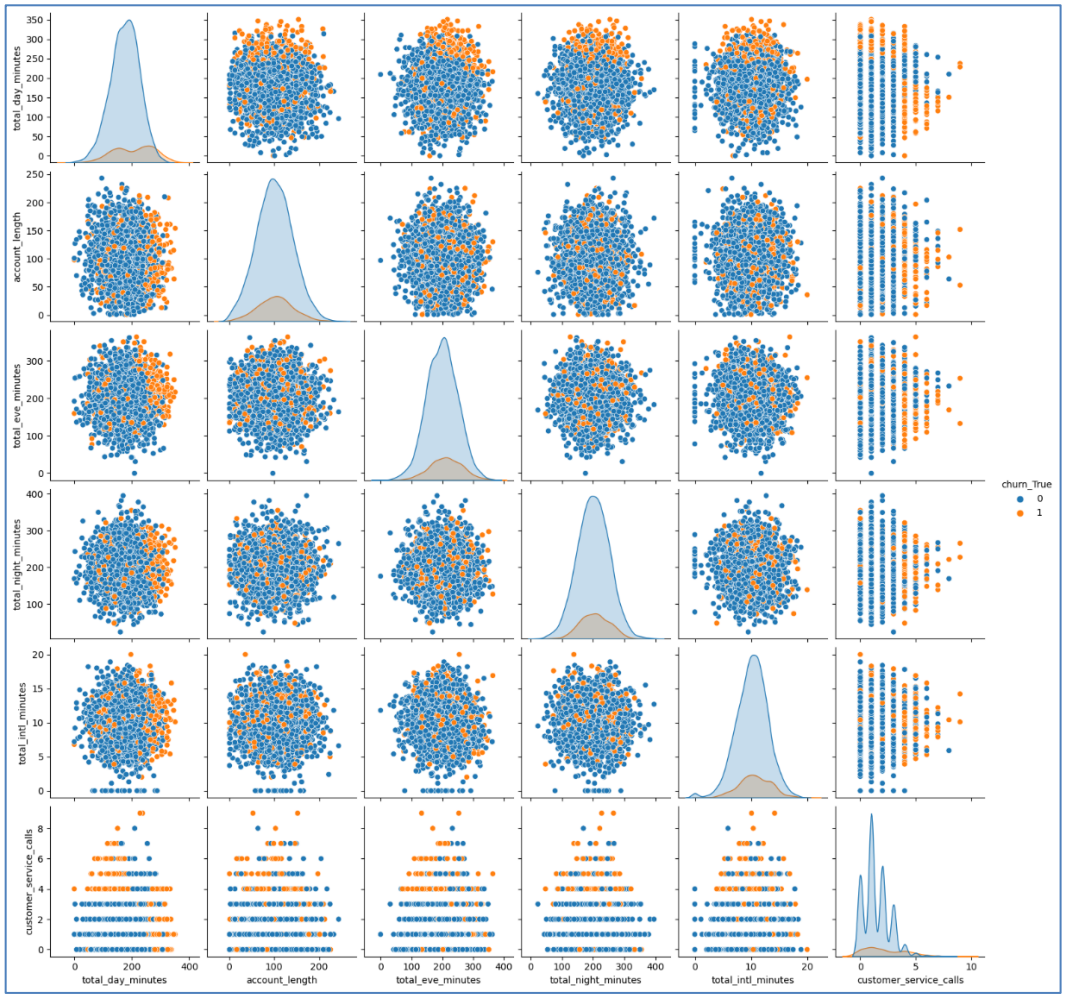




Analyzing the two bar charts reveals a significantly higher churning rate in the states of NJ, AL, AZ, WI, WA, and DC. Conversely, the churning rate is notably low in KS, OK, IA, MT, and ME.

Another observation is that, despite a higher volume of customer service calls in states like OK and RI, these issues were effectively addressed, leading to lower churning rates. In contrast, states such as AZ, AL, WI, WA, DC, and TN exhibit a higher likelihood of customer service requirements not being met, resulting in increased churning rates.

Scatter Plots:



The scatter plot above affirms that issues primarily arise with customers who engage in prolonged conversations, especially during the day—indicating challenges with heavy daytime users.

5. Classification Analysis:

Considered 70% of training data and 30% of validation data.

Normalized the dataset.

- **K-NN:** KNN Regressor is supervised ML algorithm that classifies data points based on its k nearest neighbors. This model's performance is influenced by choice of k and its distance metrics.
- **K-NN Model without 'State' Column:** K-NN model with 3 neighbors
- Accuracy: 88.9% on Validation Data, True Positive accuracy: 38%
- Removal of 'state' might have affected model sensitivity to certain patterns.

```
Confusion Matrix (Accuracy 0.8890)

       Prediction
Actual   0   1
     0 834  21
     1  90  55
```

	k	accuracy
0	1	0.868
1	2	0.875
2	3	0.889
3	4	0.887
4	5	0.892
5	6	0.891
6	7	0.891
7	8	0.883
8	9	0.888
9	10	0.882
10	11	0.889
11	12	0.884
12	13	0.881
13	14	0.878
14	15	0.888
15	16	0.877
16	17	0.877
17	18	0.875
18	19	0.876
19	20	0.876
20	21	0.877

- With the range of k 1 to 50, the best value of k is at 5.

```
Confusion Matrix (Accuracy 0.8920)

       Prediction
Actual   0   1
     0 839  16
     1  92  53
```

- **K-NN Model with 'State' Column:**
- Accuracy: 86% on Validation Data, True Positive accuracy: 15%
- Accuracy decreased after including 'state.'
- Impact: Dimensionality, feature relevance, or noise introduced by 'state.'

```
Confusion Matrix (Accuracy 0.8650)

       Prediction
Actual   0   1
     0 843  12
     1 123  22
```

- **Parsimonious Model (7 Predictors) and best K at K = 9:**
- We considered 7 predictors- **'international_plan','number_vmail_messages','voice_mail_plan','total_day_charge','total_eve_charge','total_night_charge','total_intl_charge'**

```
Confusion Matrix (Accuracy 0.9010)

       Prediction
Actual   0   1
     0 842  13
     1  86  59
```

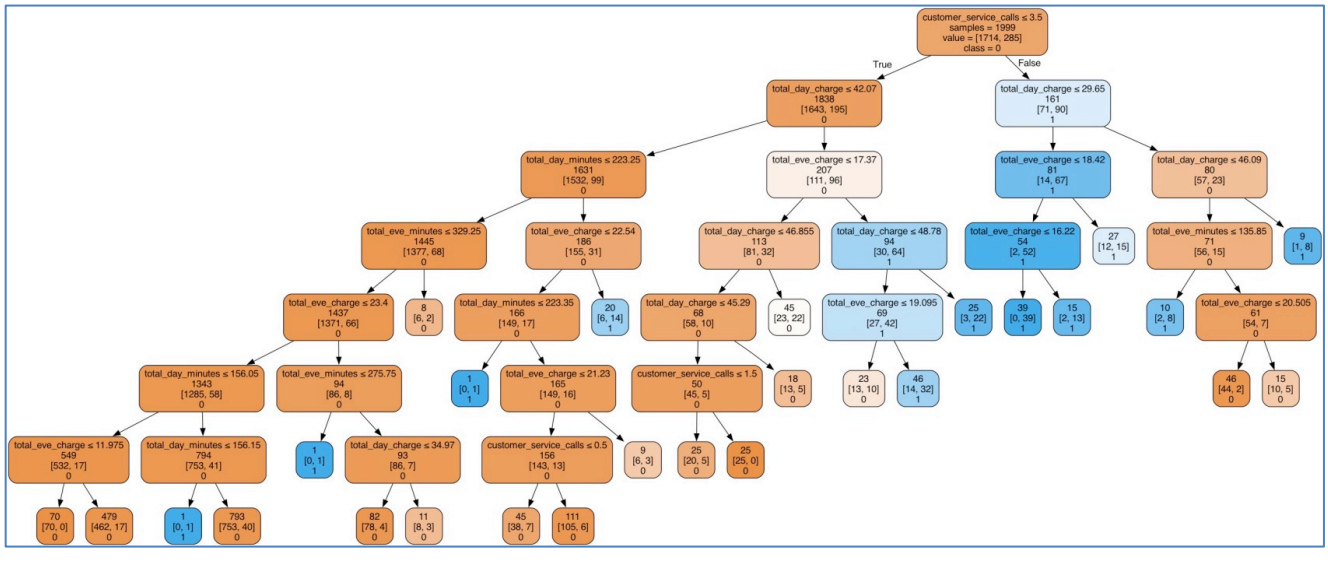
- Accuracy: 90.1% on Validation Data, True Positive accuracy: 40%
- Feature selection and model simplicity led to improved accuracy.

- **Decision Trees:**
- A decision tree is binary classification of dataset and useful for understanding feature importance. It creates tree like structure by recursively splits the dataset into subsets based on the most significant feature at each node.
- Created the tree with maxdepth = 7 and min_samples_split =50
- An accuracy of 92.7% on validation data predicted for a naive model with all the predictors.

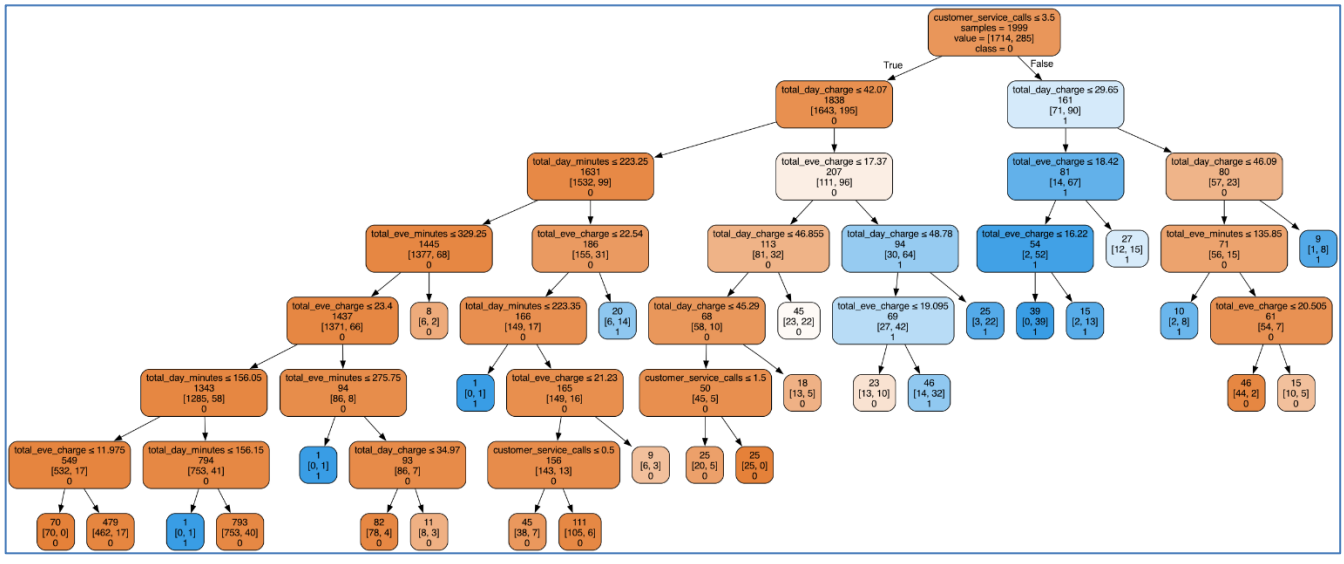
```
classificationSummary(valid_y, classTree1.predict(valid_X))

Confusion Matrix (Accuracy 0.9273)

       Prediction
Actual    0    1
     0 1099   37
     1   60  138
```



- The parsimonious model with 5 Top predictors gave an overall accuracy of 88.68% on the validation data.



- Used **grid search method** to enhance accuracy even further.

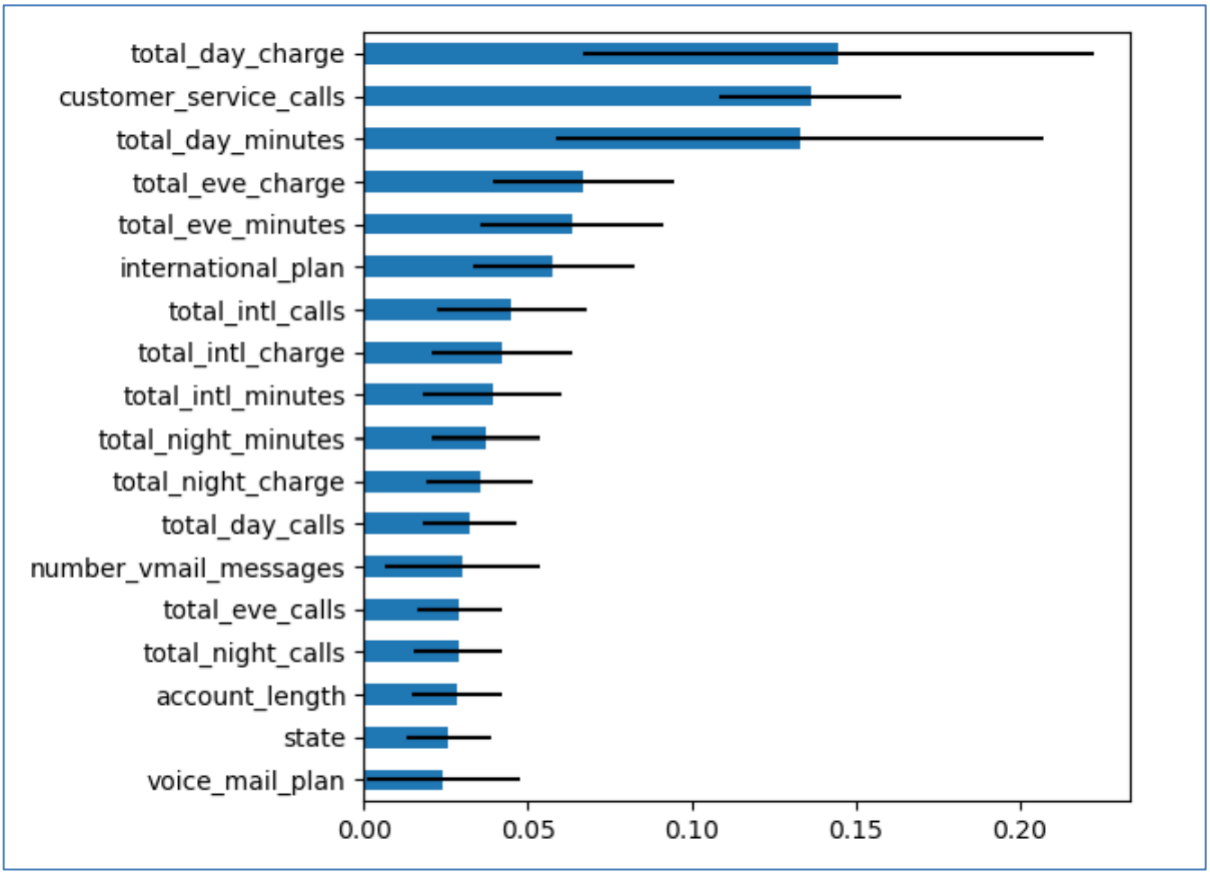
```
classificationSummary(valid_y, bestClassTree.predict(valid_X))

Confusion Matrix (Accuracy 0.9333)

       Prediction
Actual    0    1
     0 1103   33
     1   56  142
```

Grid search parameters gave an overall accuracy of 93.3% on validation data.

- **Random Forest Trees:**
- Random forest is a powerful machine learning method that boosts the accuracy of predictions by using a combination of multiple classification trees. It trains each decision tree based on random subset of data and then combines all predictions from those trees. The collective predictions of the individual trees provide highly accurate and reliable predictions for classification or regression problems.
- Achieved a validation accuracy of 95%, with a True Positive accuracy of 72%.
- Additionally, identified and utilized key features to construct a parsimonious model. Important features are **'total_day_charge', customer_service_calls', 'total_day_minutes', 'total_eve_charge', 'total_eve_minutes', 'international_plan'.**




```
classificationSummary(valid_y, rf.predict(valid_X))

Confusion Matrix (Accuracy 0.9505)

       Prediction
Actual    0    1
     0 1124   12
     1   54  144
```

- Random forest gave an accuracy of 95% on the validation data.
- **Logistic Regression:** Base model with all the predictors gave an accuracy of 85% on the validation data and True positive accuracy as 14%.

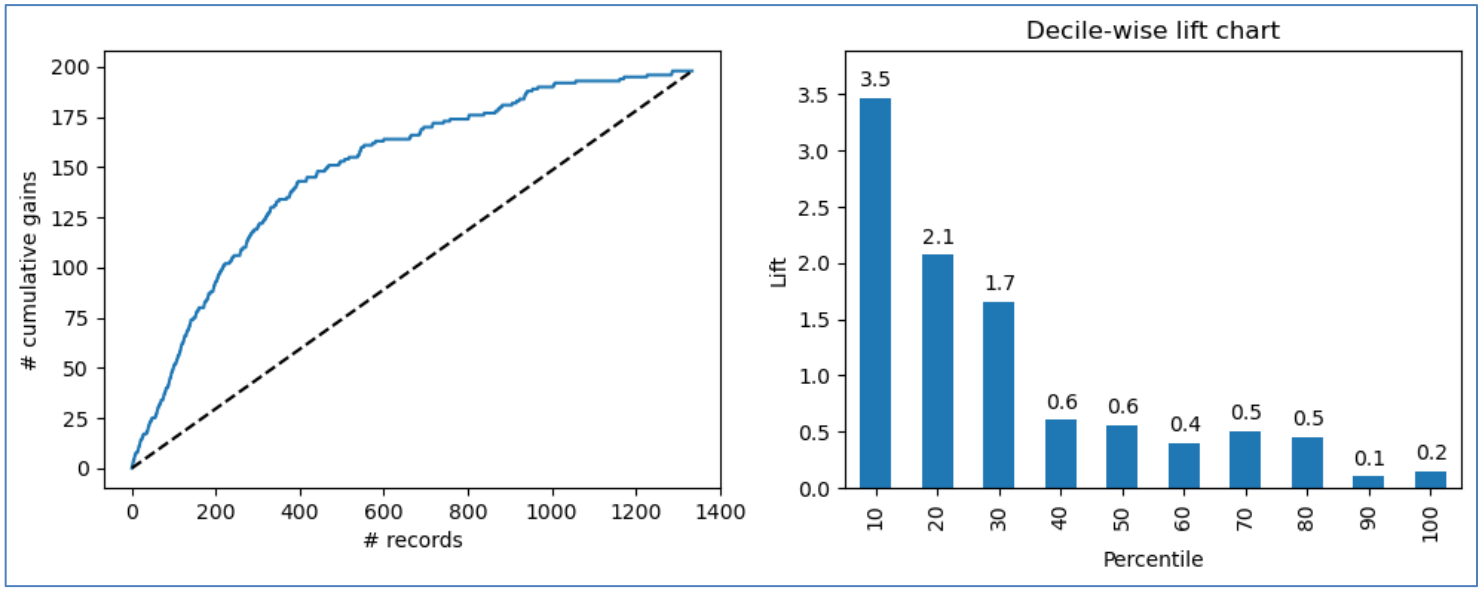
```
classificationSummary(valid_y, logit_reg_pred) #

Confusion Matrix (Accuracy 0.8501)

       Prediction
Actual    0    1
     0 1089   47
     1  153   45
```

- Naive logistic regression model gave an overall accuracy of 85% on the validation data.

- **Gains chart and Decile-wise lift chart:**



- **Logistic regression model with the top predictors identified through the Tree Classifier:**

```
logit_reg_pred = logit_reg.predict(valid_X)
classificationSummary(valid_y, logit_reg_pred) #testing accuracy on va
```

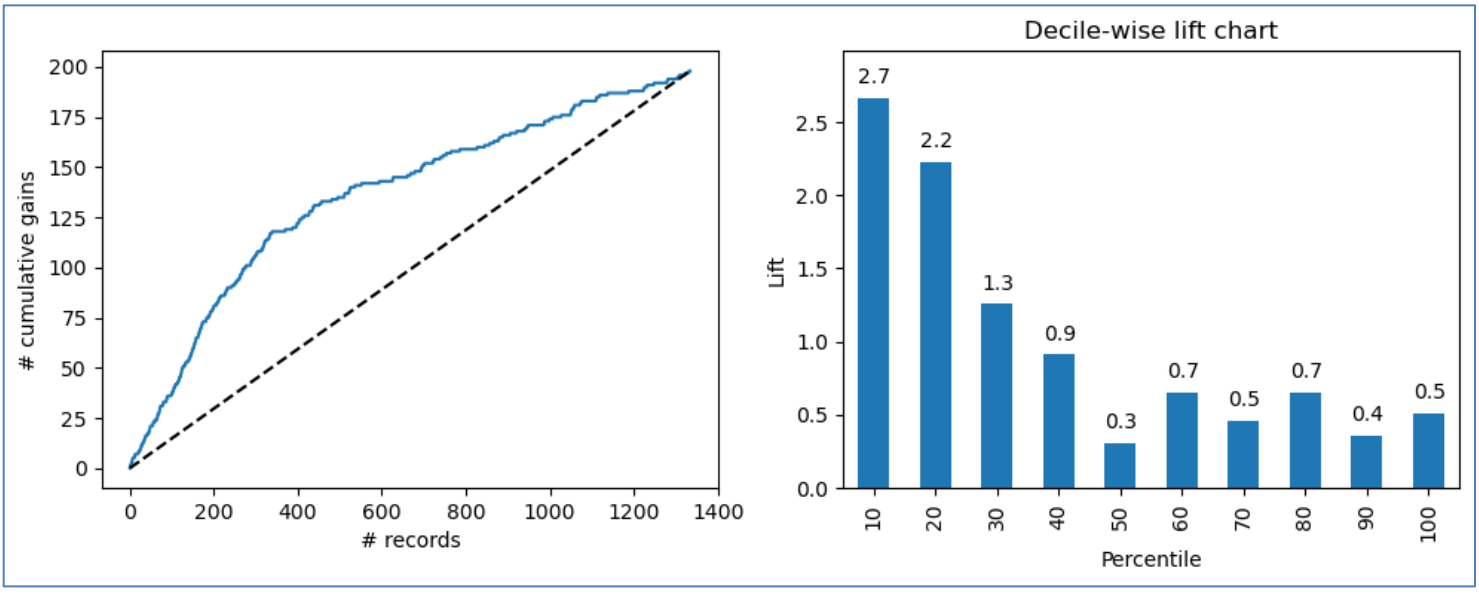
```
intercept  -6.3500427124198735
       total_day_charge  customer_service_calls  total_day_minutes  \
coeff          0.004786                0.498938           0.012628

       total_eve_charge  total_eve_minutes
coeff         -0.006303           0.005624

AIC 1301.2430586422943
Confusion Matrix (Accuracy 0.8463)

       Prediction
Actual    0    1
     0 1120   16
     1  189    9
```

- AIC: 1301
- Achieved 84% accuracy on the validation data.
- True positive accuracy reached 4%.



6. Conclusion:

Comparison of various Models

Model	Description	Overall Accuracy	True Negative	True Positive
KNN	All Variable(k=3)	88%	97%	38%
	All Variable (k=5)	89%	98%	37%
	All Variable (Including State)(k=3)	86%	99%	15%
	Selected variables (Different Combinations) (k=3)	89%	97%	43%
	Selected Variable 7 (k=9)	90%	98%	41%
Decision tree	Base Model(depth 7)	92%	97%	70%
	Grid Search	93%	97%	72%
	Random Forest (BEST)	**95%**	**98%**	**72%**
	Boosting	**95%**	**98%**	**74%**
	Parsimonious Model (Importance feature)	88%	96%	44%
Logit	Base Model	85%	96%	23%
	Model with important predictors	84%	98%	5%

- Best Overall Performers: Boosting and Random Forest achieve the highest accuracy among the models. In the context of a telecommunications business, considering not only overall accuracy but also true positive rates is crucial. Both Boosting and Random Forest emerge as strong candidates due to their high accuracy and a well-balanced trade-off between sensitivity and specificity.
- Business Perspective: However, from a business perspective, the goal is often to differentiate between types of customers. In this scenario, regression trees may offer a better solution as they can provide more

detailed insights by creating splits that enhance our understanding of different customer segments.

- Additionally, logit can be valuable in predicting binary outcomes, making it useful for understanding the likelihood of customer churn. Logit models provide probabilities, enabling us to assess the probability of a customer churning based on specific predictor variables. This information can be valuable for targeted interventions and personalized retention strategies.

Recommendations

- Launch **daytime unlimited** calling plans for heavy users to mitigate churning caused by high charges associated with extended talk time.
- Promote off-peak calling with discounts, flexible packages, and awareness campaigns for optimized network use.
- Utilize predictive modeling to create personalized offers for at-risk customers, addressing their specific needs and preferences to increase engagement and loyalty.
- Enhance customer service in states with unmet needs through agent training and implement a robust feedback system for improved issue resolution.
- Strengthen the company's presence on social media platforms to engage with customers, address concerns, and build a positive online community. Social media can be a powerful tool for customer communication and brand advocacy.