Customer Churn Prediction for a Telecom Company

Group4 :
Jasvitha Buggana
Krupa Shah
Manan Upadhyay
Preksha Shah
Shivani Agrawal

Agenda

1	Problem Context and Research Questions
2	Dataset Description
3	Pre-Processing Done & Initial Exploration
4	What we have done
5	Analysis and Results
6	Comparison of different models
7	Conclusion & Recommendations

Problem Context & Research Questions

Current Churn

- 14 % churn rate

Company Services

- Vast Coverage (all states)
- Customer Support

Features of Talk Plans

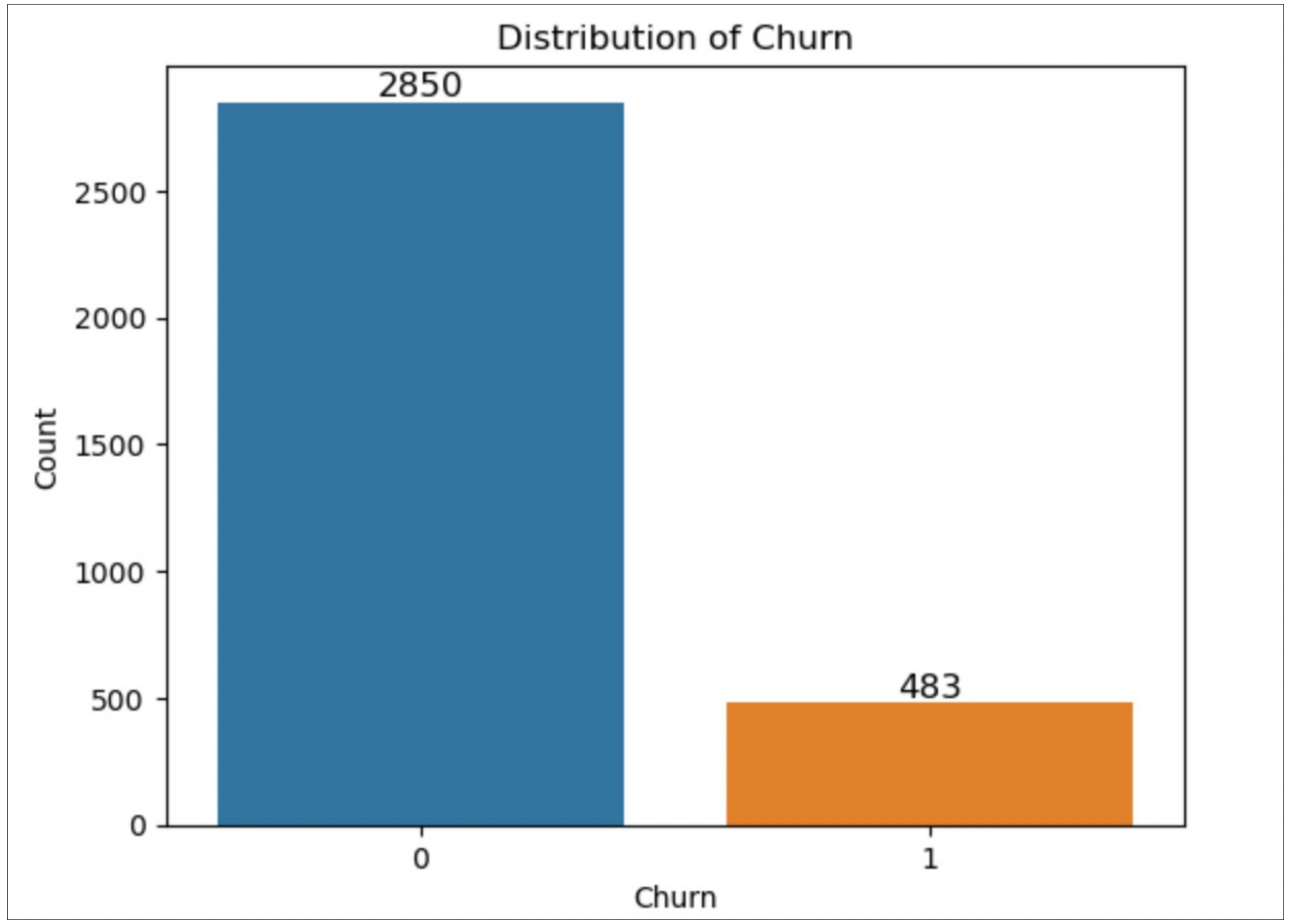
- International
- Voicemails, Day, Night,

Problems to Identify
Are there any particular plans like day, eve, and night which are facing higher churns?
Are there any states where the churn rate is higher?
How is our customer service team performing?
Is there any specific geography where the customer team is not doing well, and hence customers are churning?
What is the main pain point?

Data Set Description

The Data set consists of 3333 rows and 21 Columns

Source:
[https://platform.stratascratch.com/data-pr
s/customer-churn-prediction](https://platform.stratascratch.com/data-pr)




```
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   int64
dtypes: float64(8), int64(9), object(4)
memory usage: 546.9+ KB
```

Pre-processing & Initial Exploration

Data Type Conversion to Category

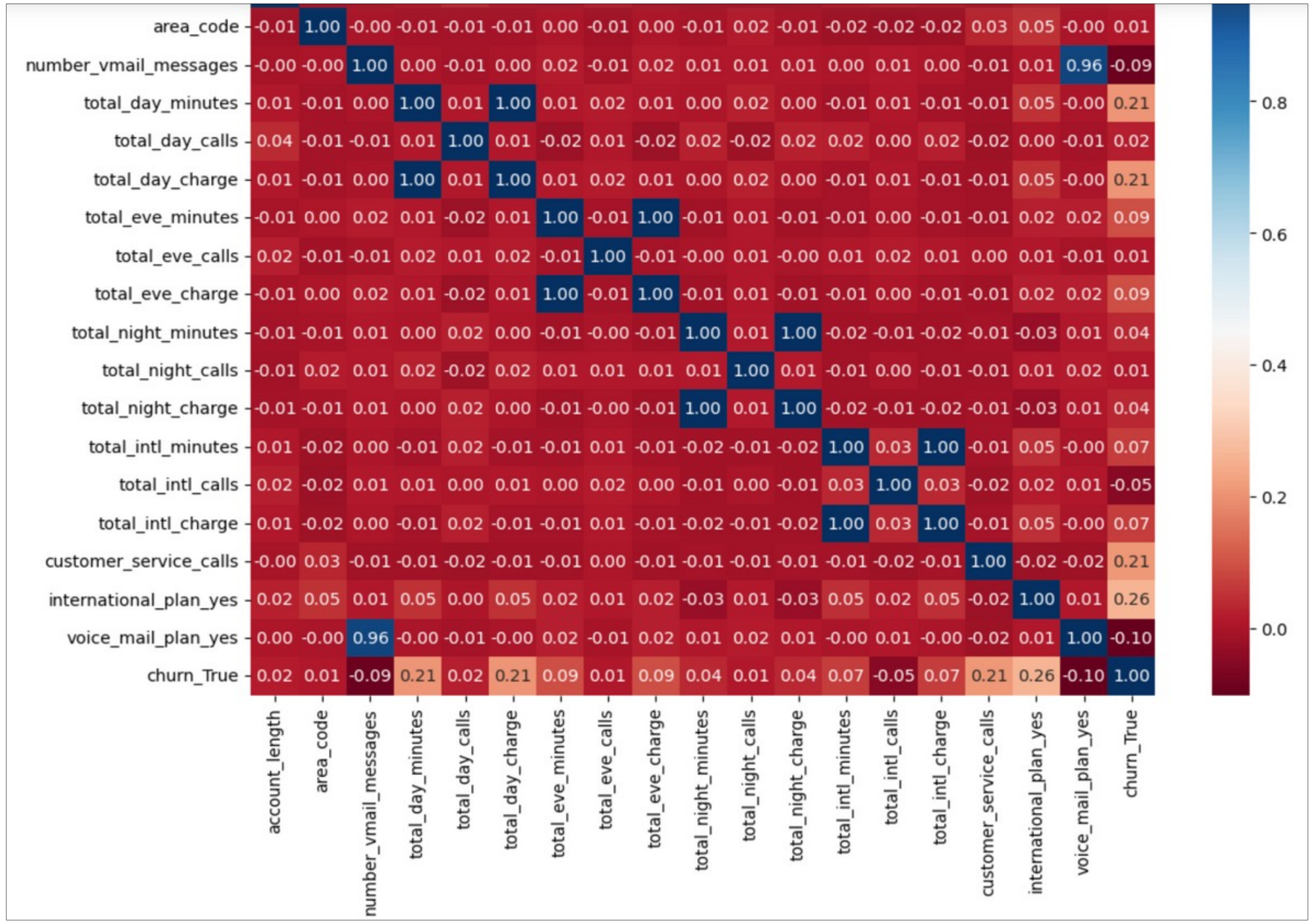
- Changed Outcome Variable: Churn from Text to 0/1, State, International_plan, Voice_mail_plan
- Numerical Data : int, float
- Non-Numerical: String / Object

Encoding

- Coded State Name to a number, E.g.: CA became 4

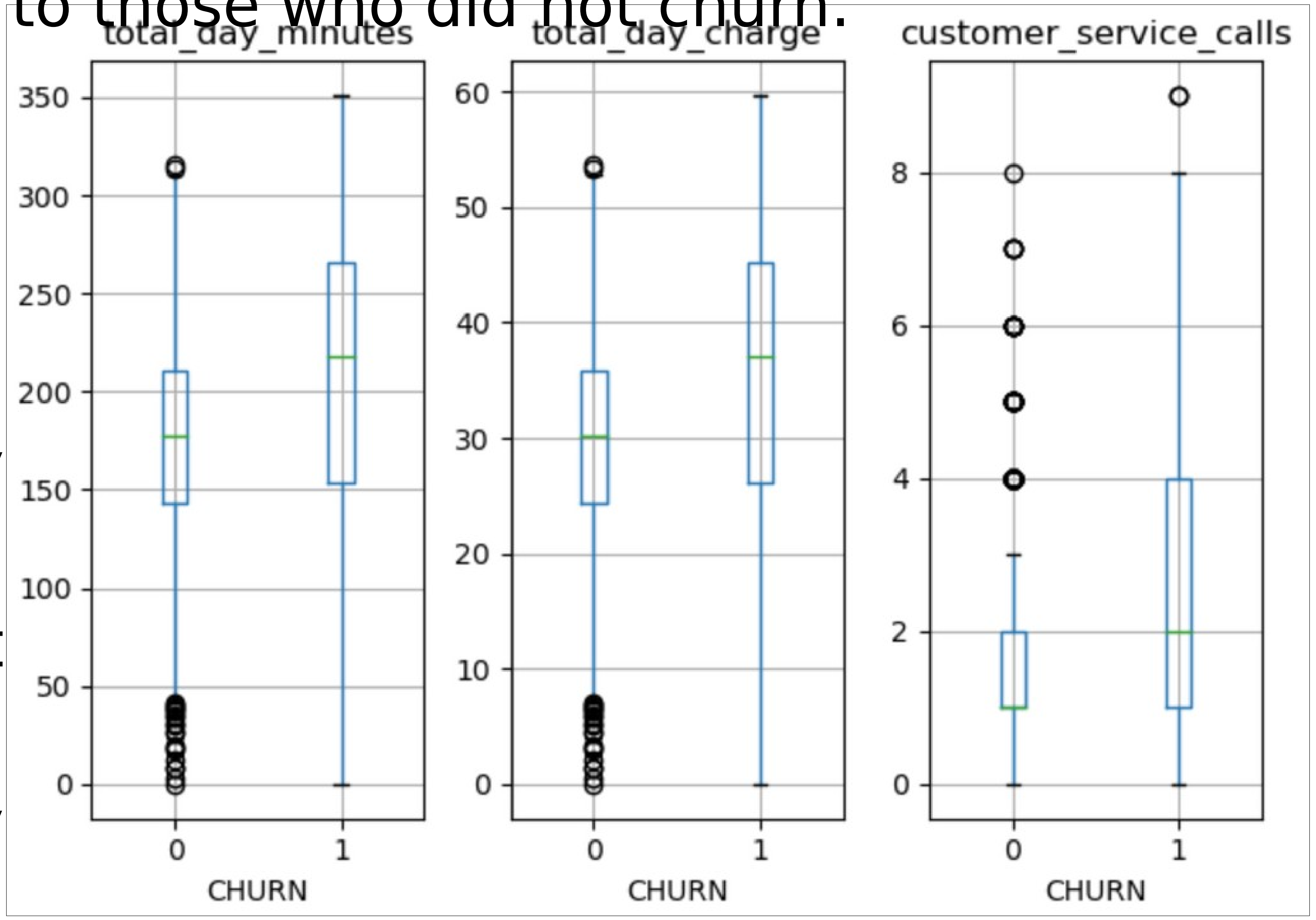
Name Change

- Stripped space from the original variable names.
- Account Length became account_length

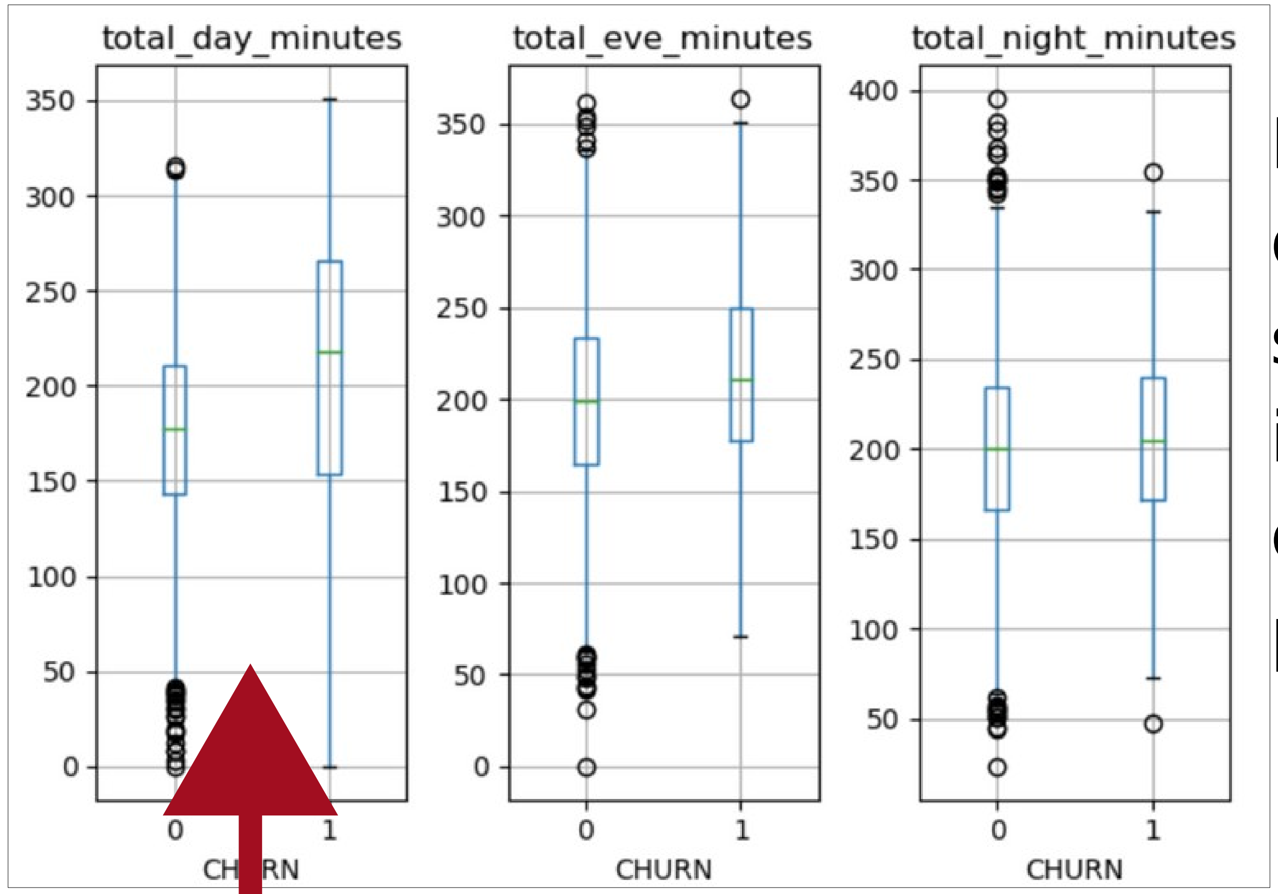


VISUALIZATION(Preprocessing and Initial Analysis)

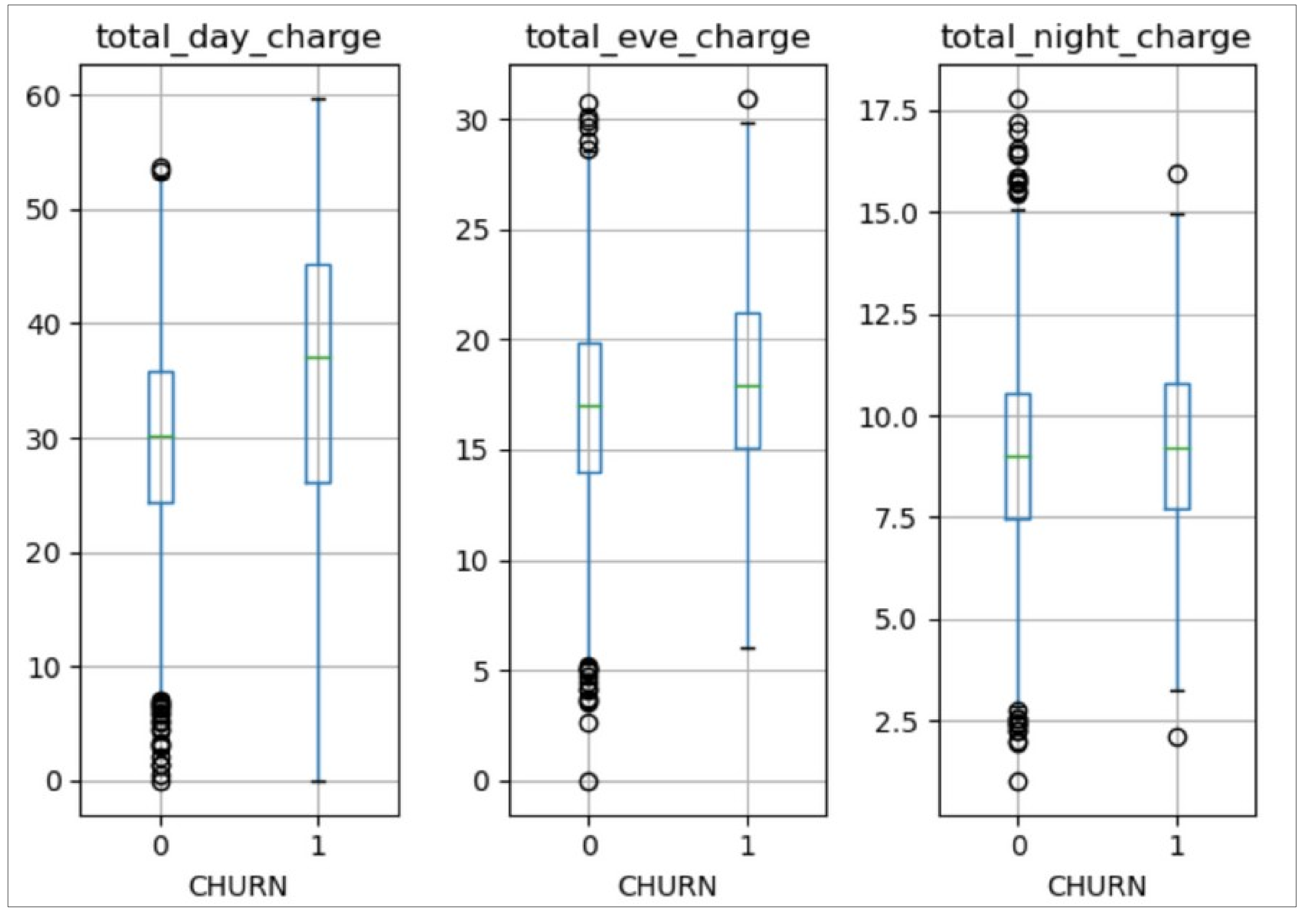
- The Heat-map does not provide relevant information about the collinearity so we move towards Box-plot Analysis.
- Those who churned engaged in more extensive conversations and made relatively higher payments compared to those who did not churn.



- Customers who churned experienced significantly higher volumes of customer service calls, indicating that the issues raised during these calls were not adequately addressed.

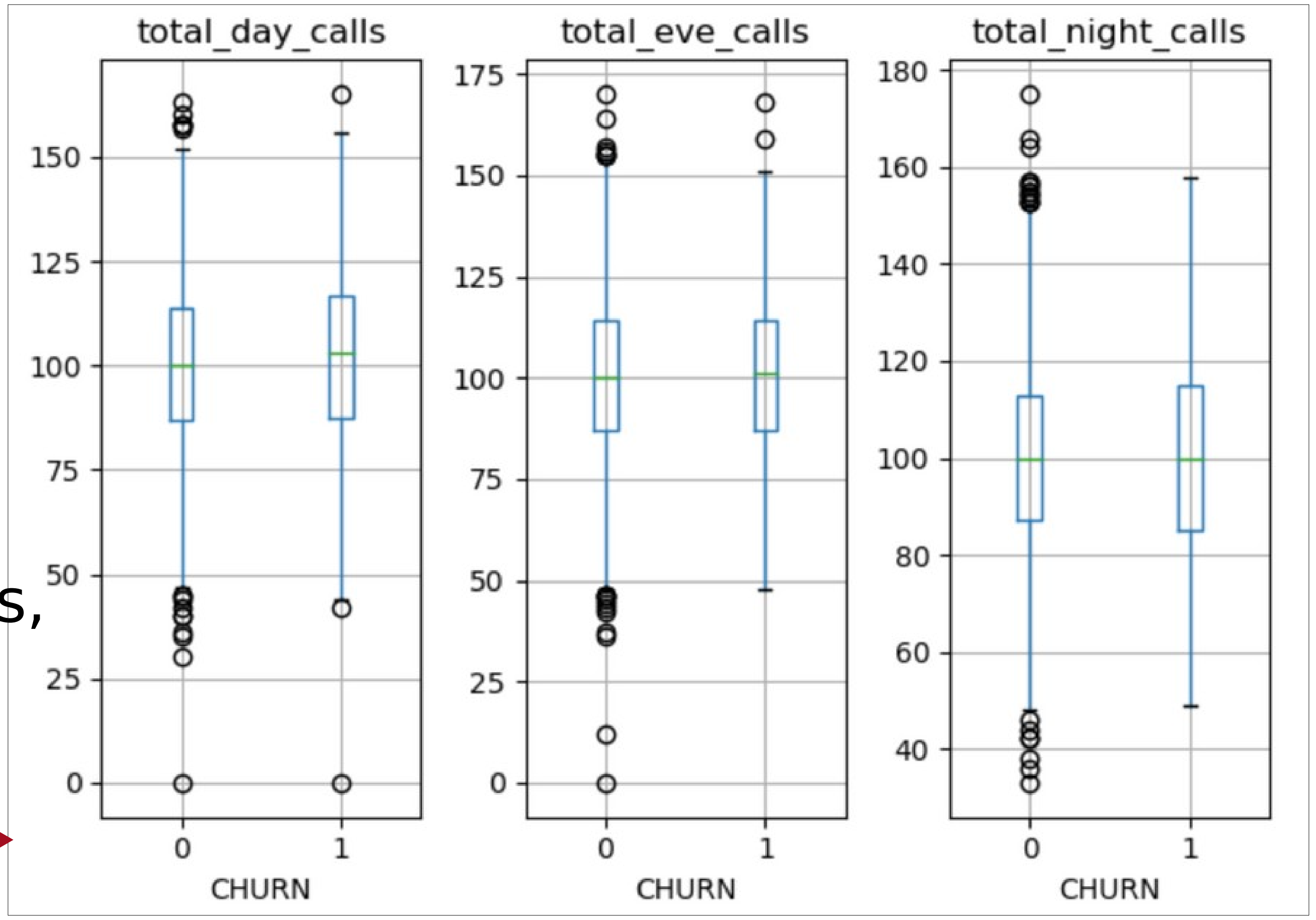


Daytime charges were higher despite an equal call count, suggesting that churned individuals had longer conversations and made greater payments.

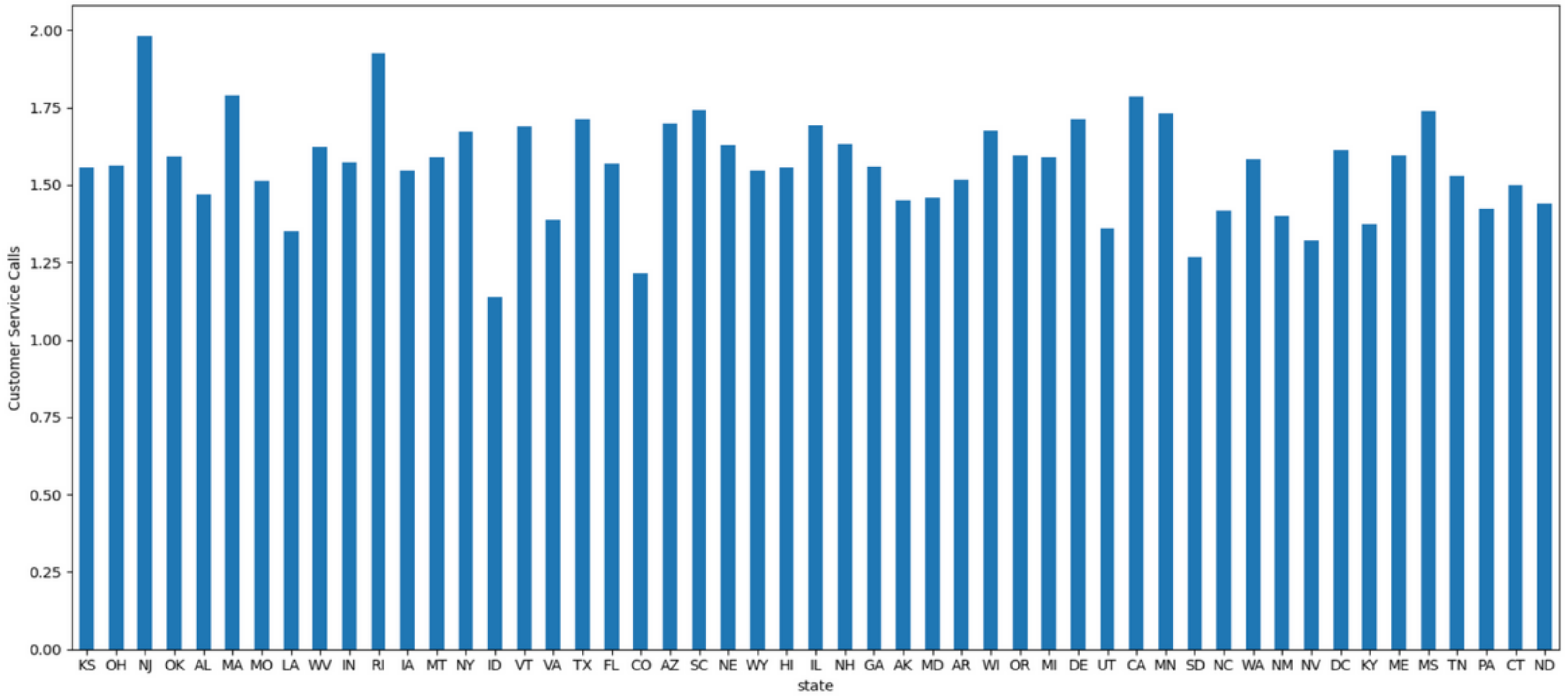


Daytime witnessed the longest call durations throughout the day for the people who churned.

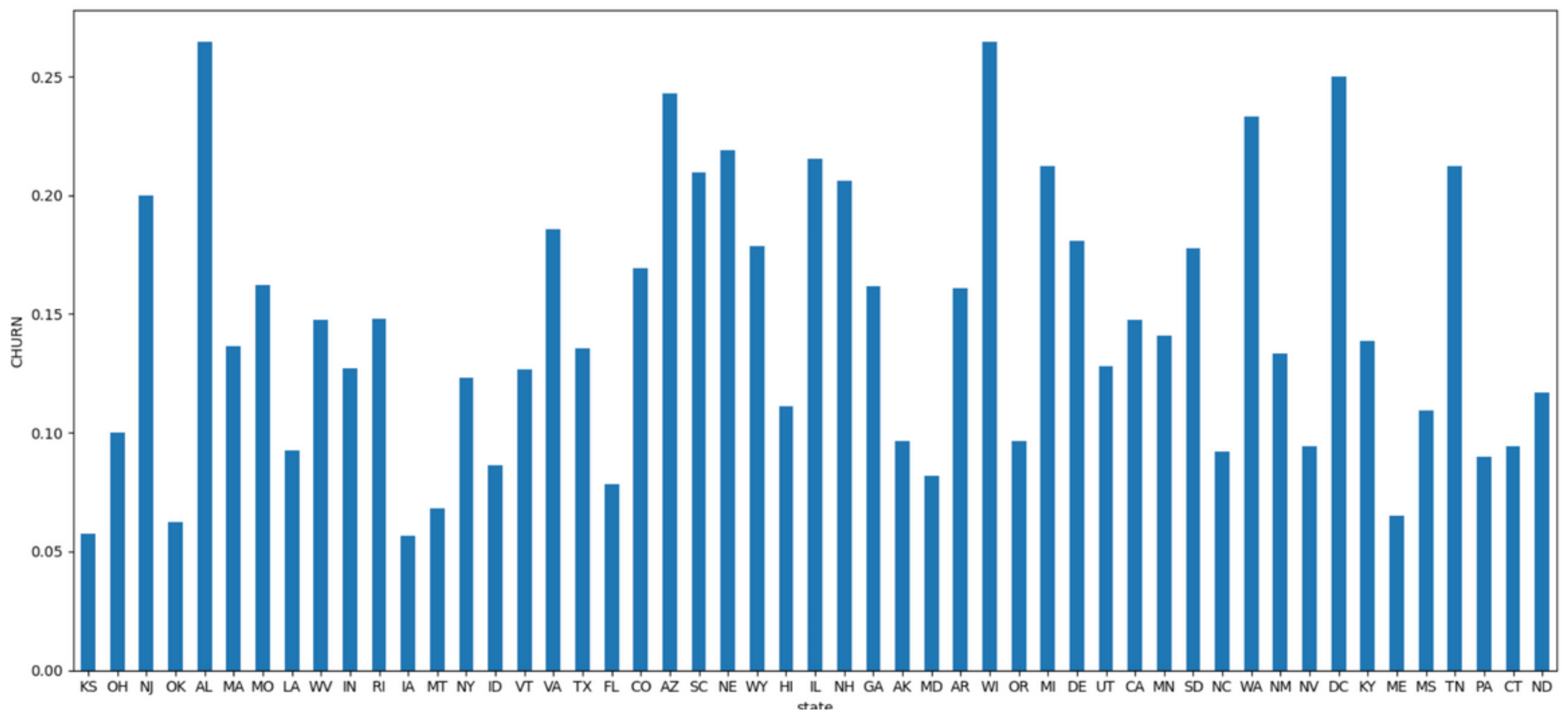
Despite consistent call counts for both churned and non-churned customers, daytime callers, especially among those who churned, had lengthier conversations.

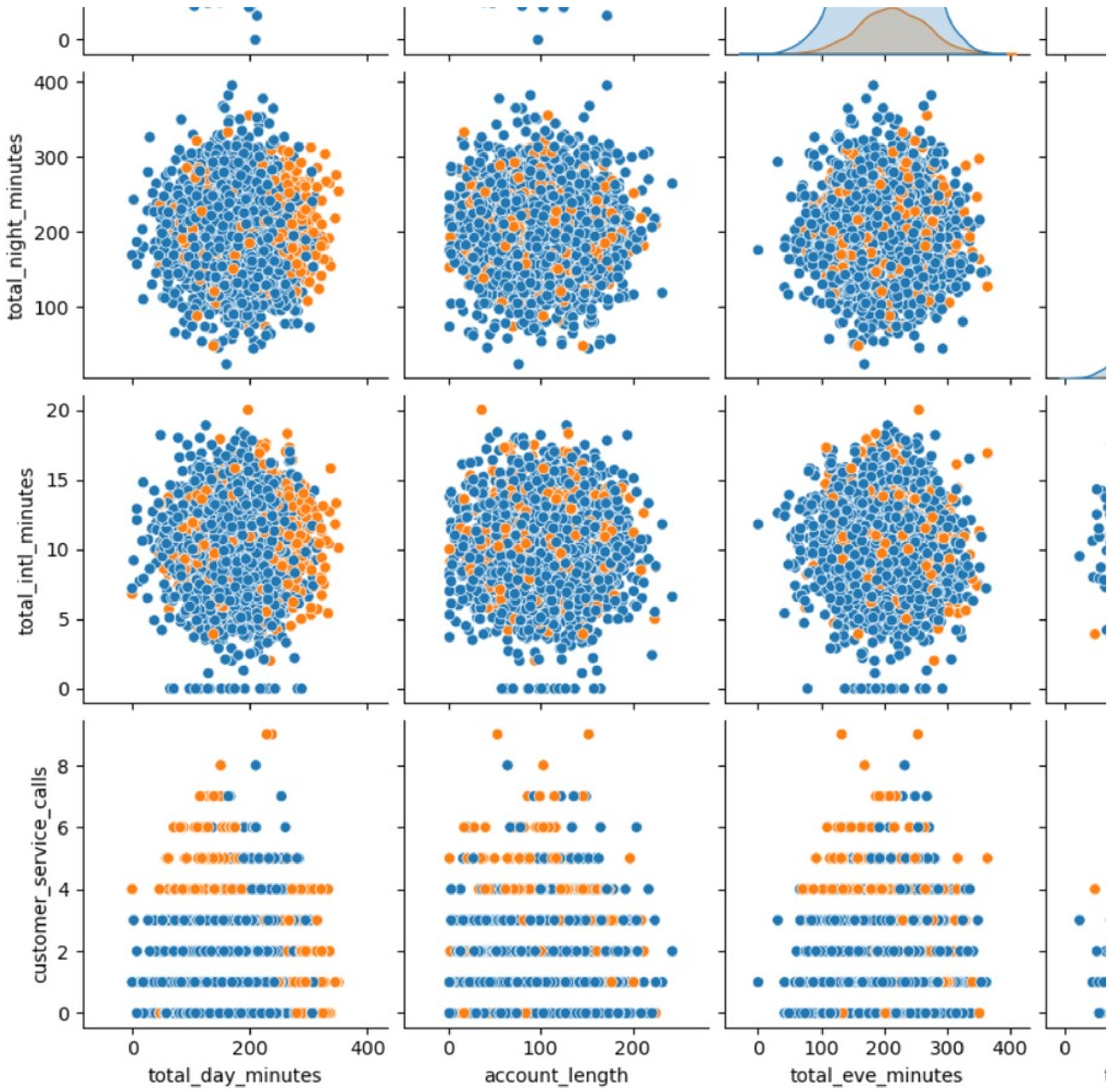


Other telecom companies may offer unlimited plans for extended calls at a fixed price, while our company charges based on minutes.



High churning: NJ, AL, AZ, WI, WA, DC.
Low churning: KS, OK, IA, MT, ME.
Resolved issues in OK and RI led to low churning despite more calls. Likely unmet needs in AZ, AL, WI, WA, DC, TN led to higher churning.





The scatter plot confirms issues predominantly arise with customers who engage in extended daytime conversations, particularly heavy daytime users.

What Have We Done?

Data Set	KNN	Base Model
		Second Model with Best K
		Third KNN model with a few predictors only (parsimonious KNN Model)
	Decision Trees	Base Model
		Another Model using Grid Search
		Random Forest **(Best)**
		Boosting
	Logistic Regression	Base model
		Model with important predictors only

Analysis & Results K-NN model Performance

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

K-NN Model without 'State' Column:

- 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)

Actual \ Prediction	0	1
0	834	21
1	90	55

K-NN Model with 'State' Column:

- Accuracy: 86% on Validation Data, True Positive accuracy: 15%
- Accuracy decrease after including 'state.'
- **Impact**: Dimensionality, feature relevance, or noise introduced by 'state.'

Confusion Matrix (Accuracy 0.8650)

Actual \ Prediction	0	1
0	843	12
1	123	22

Parsimonious Model (7 Predictors) and best K at K = 9:

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

Confusion Matrix (Accuracy 0.9010)

Actual \ Prediction	0	1
0	842	13
1	86	59

Analysis & Results Decision Tree Performance

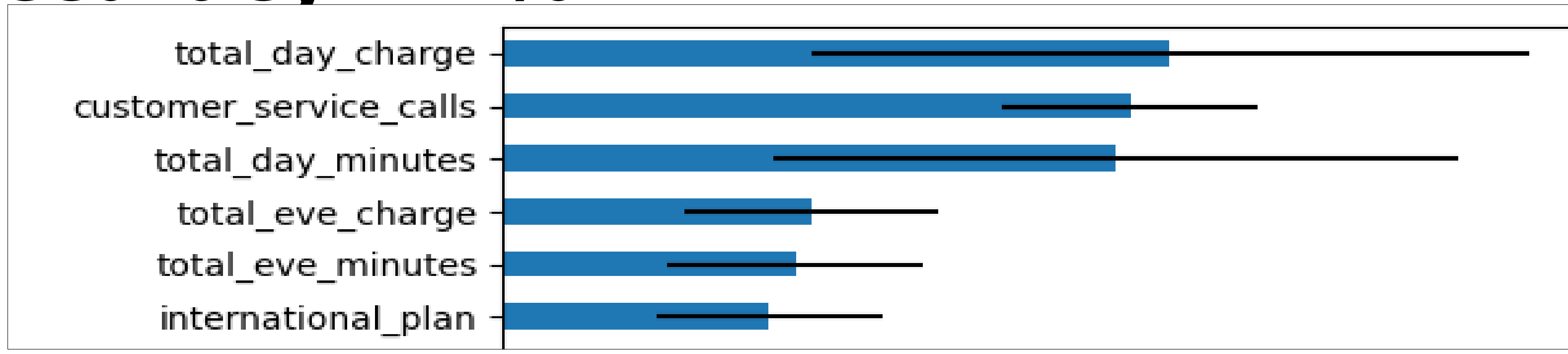
```
classificationSummary(valid_y, rf.predict(valid_X))

Confusion Matrix (Accuracy 0.9505)

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

Mentioning about the best and parsimonious model.

Random Forest

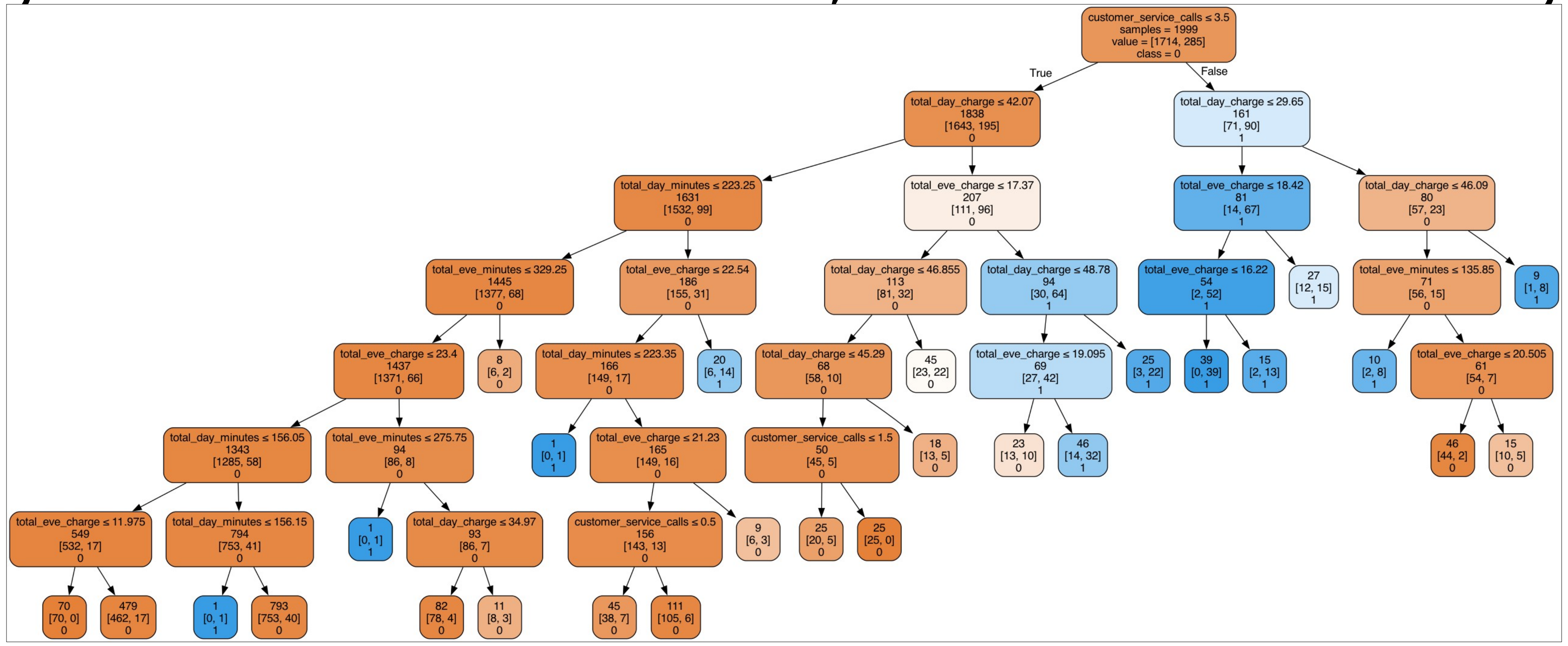
- Accuracy: 95% on Validation Data, **True Positive accuracy: 72%**
- Also got to know the important featured

which we used to make the parsimonious model.



Parsimonious Model (Top 5 Predictors):

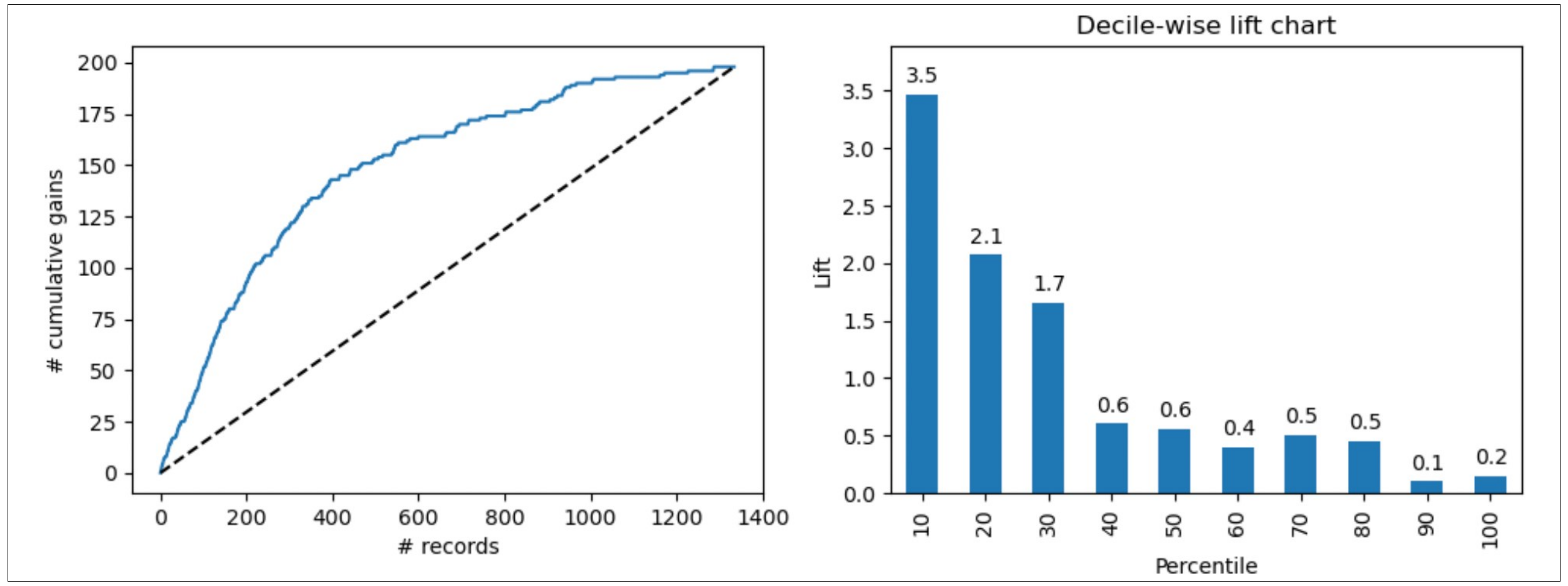
- Accuracy: 88% on Validation Data, True Positive accuracy: 44%



Analysis & Results Logistic Reg. Performance

Base Model with all predictors

- AIC: 1392
- Accuracy 85% on Validation Data
- True Positive accuracy: 14%



Parsimonious Model (7 Predictors) and best K at K = 9:

- AIC: 1301
- Accuracy: 84% on Validation Data
- True Positive accuracy: 4%

```
Confusion Matrix (Accuracy 0.8463)

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

Comparison of All Models

Model	Description	Overall_Accuracy	Sensitivity	Specificity
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%

Recommendation & Conclusion

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.

Enhance customer service in states with unmet needs through agent training and implement a robust feedback system for improved issue resolution.