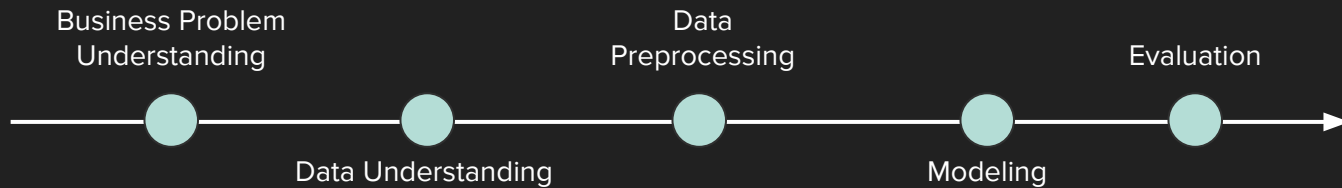




***PREVENTING CHURN,***  
*Keeping Customers Happy*



## Data Understanding

---

*Telecom Dataset from  
Kaggle*

**kaggle**

*3333 customers from all  
50 states*

*Length, frequency, & cost of  
calls, subscription to  
international plan, voicemail  
plan*

# CHURN

**5 to 25 more expensive** to acquire new customers, compared to retaining current customers

Increasing retention rates by 5% can increase **profit rates by 25-95%**

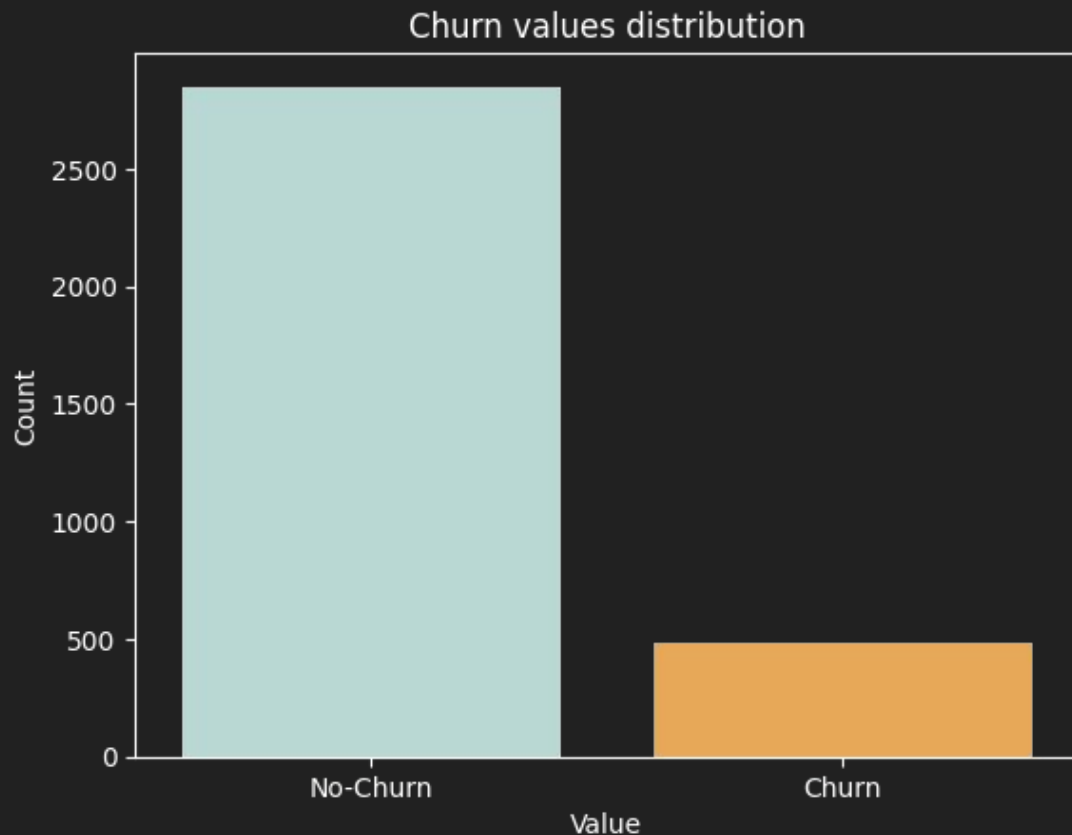
# 86%

*of data is no-churn  
customers*

# 14%

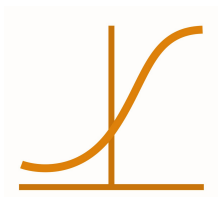
*of the data is churn  
customers*

*SMOTE (Synthetic Minority  
Over-sampling Technique)*

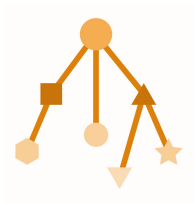


# MODELING

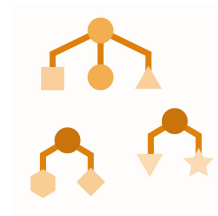
Logistic regression



Decision tree



Random forest



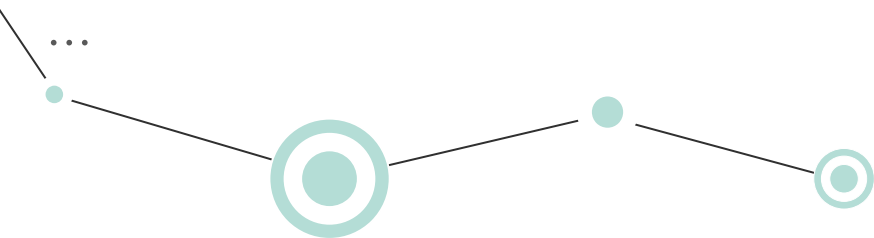
# EVALUATION METRICS

*PRECISION*

*RECALL*

*F1 SCORE*

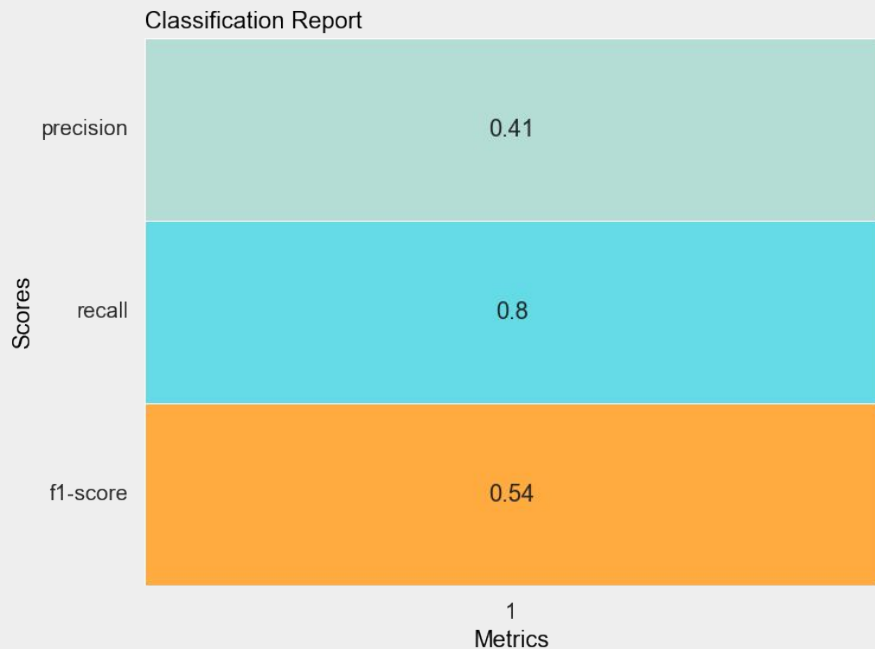
*ROC-AUC*





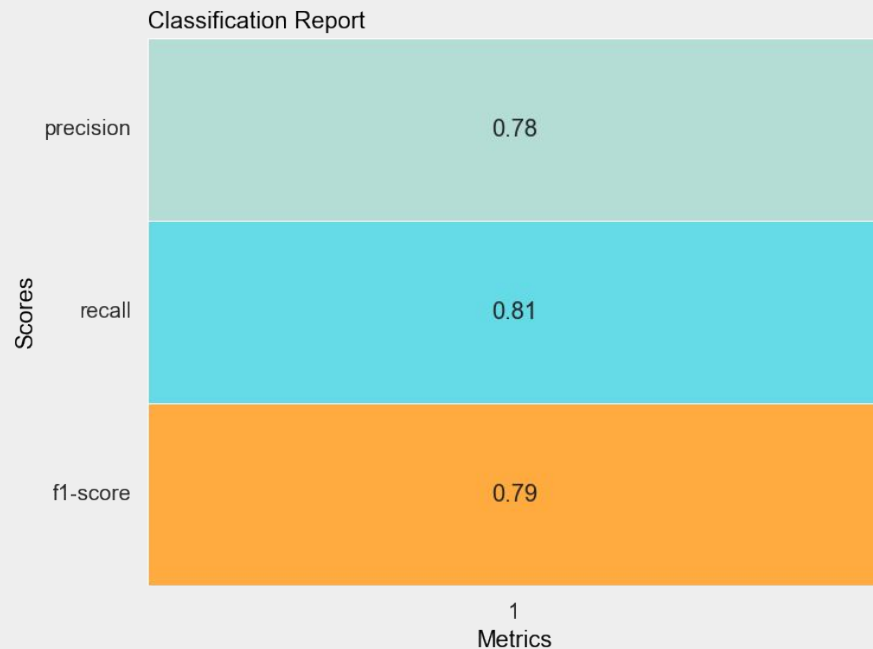
# Logistic Regression

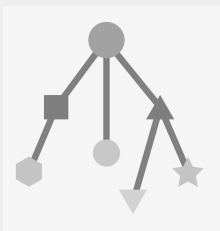
## RAW SCORES



## FINAL SCORES

(Cross-validated, hyper-tuned)

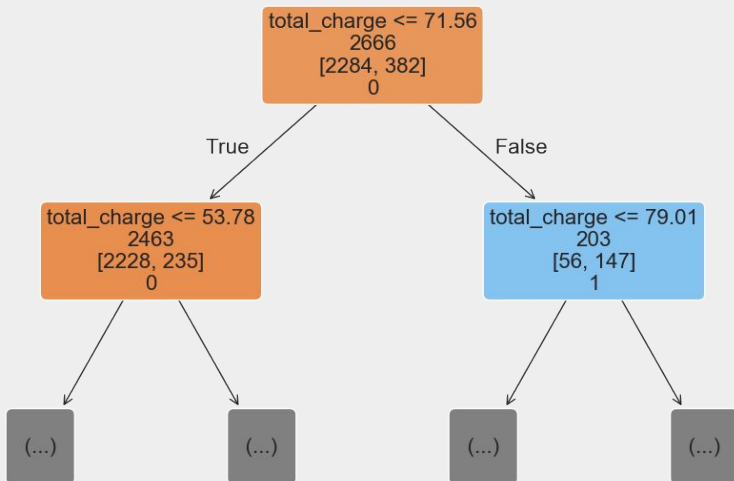




# Decision Tree

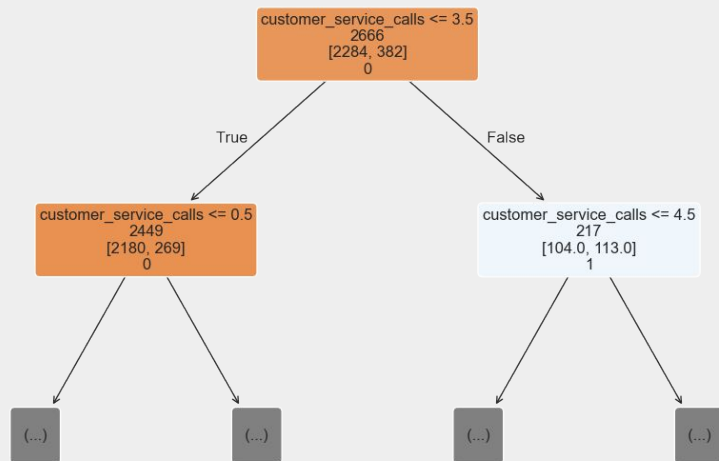
*“customers with a total-charge **less** than 71.56 are likely to **not** churn”*

Decision Tree for total\_charge



*“customers with a customer-service call count **less** than 3.5 are likely to **not** churn”*

Decision Tree for customer\_service\_calls

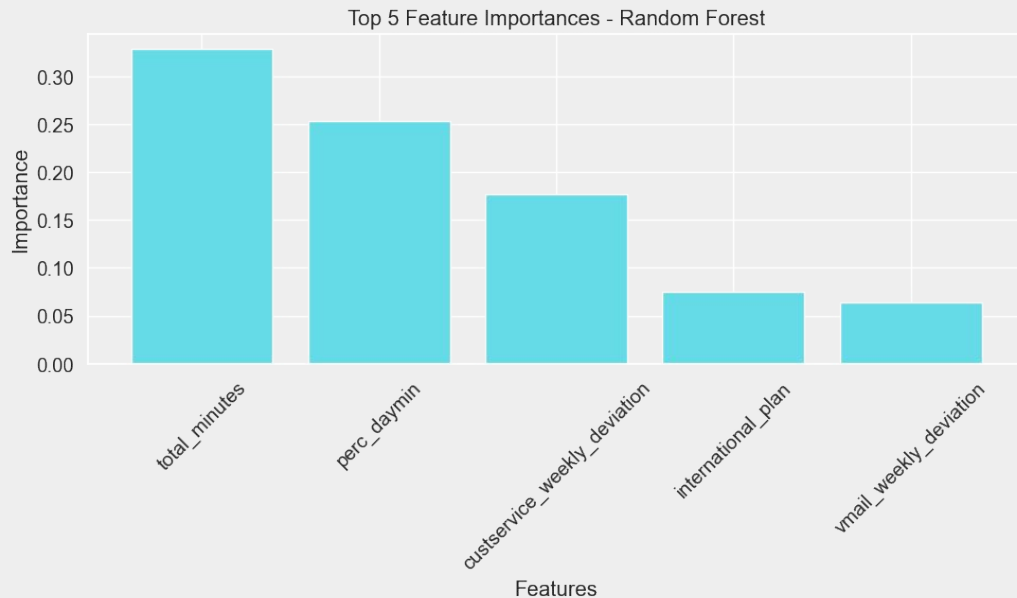


# Feature Selection



## Random Forest: Feature Selection

- Less sensitive to multicollinearity, captures non-linear relationships
- Ensembling method: combines predictions of multiple base models (decision trees)
- Measures reduction in purity across all decision trees in the forest



## Logistic Regression

- Coefficient values
- Impact scale of 0-1

customer  
service: **0.77**

total  
minutes: **0.76**

international  
plan: **0.68**

voice-mail  
plan: **-0.36**



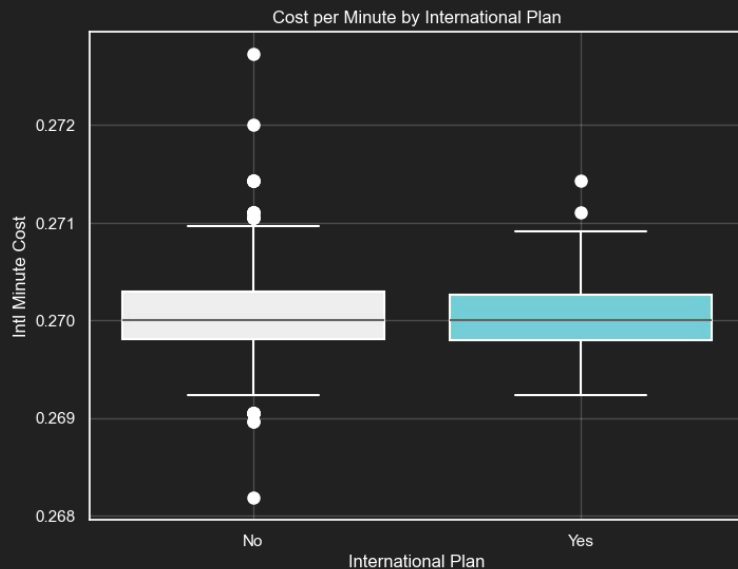
# INTERNATIONAL PLAN

#3

Most important feature in final  
logistic regression model

0.68 coefficient

(subscribers likely to **churn**)



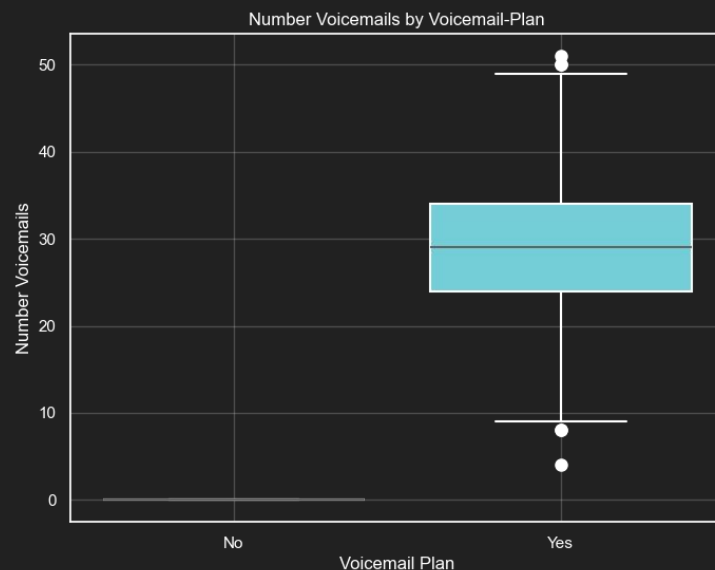
# VOICEMAIL PLAN

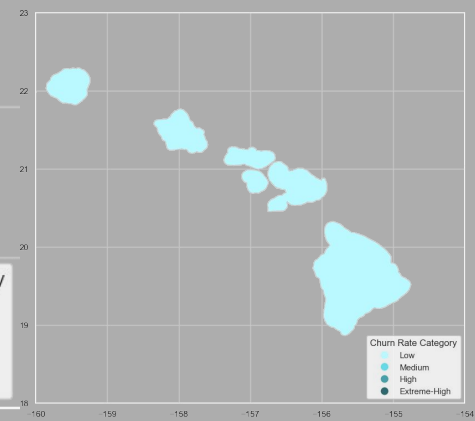
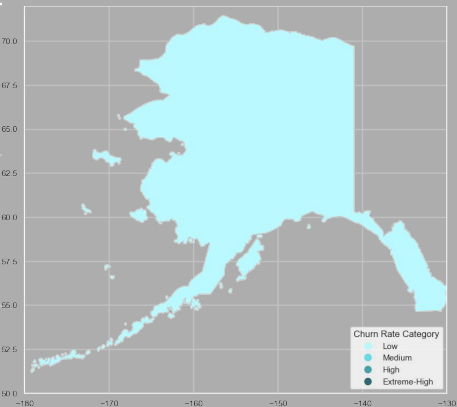
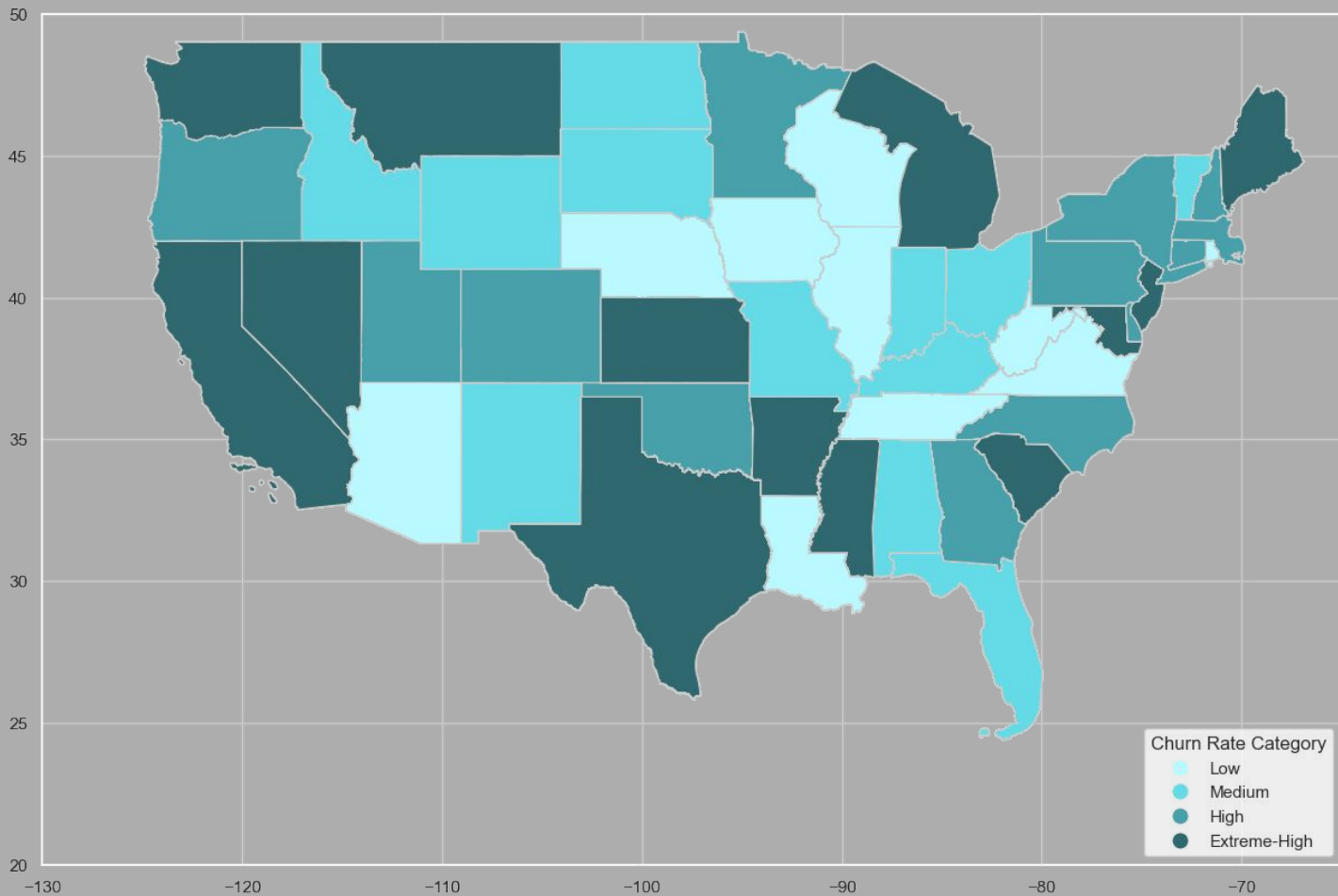
#4

Most important feature in final  
logistic regression model

-0.36 coefficient

(subscribers likely to **not churn**)





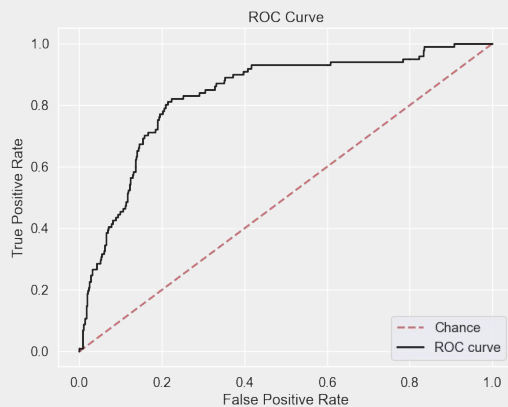


# BEST MODEL: LOGISTIC REGRESSION

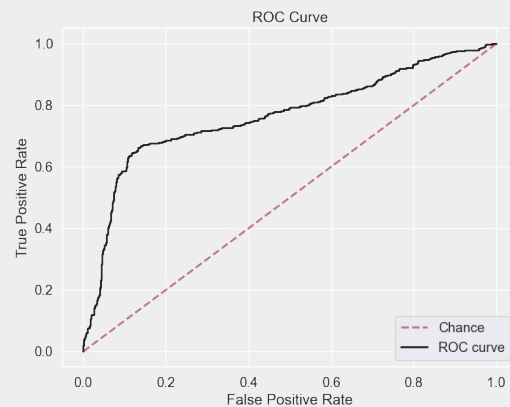




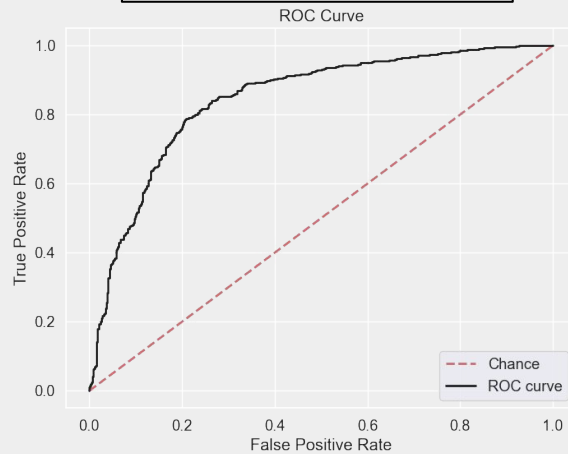
RAW  
ROC-AUC: 0.83



MID-TUNING  
ROC-AUC: 0.85



FINAL  
ROC-AUC: 0.85



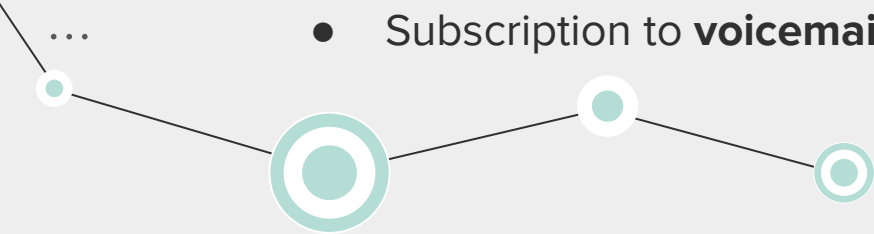
# CONCLUSIONS



Key behaviors predictive of churn:

- 2+ **customer service calls**
- Increased **number of call-time minutes**, particularly during the **daytime** (most likely related to *increased charges*)
- Subscription to **international plan**
- Residing in a **high churn-rate state**

Key behaviors prediction of *no* churn:

- Subscription to **voicemail plan**
- 

# NEXT STEPS

---

## Duration of Call & Daytime Calling

PRICE-PER-MINUTE (by time of day)

DAY: 0.17

EVENING: 0.085

NIGHT: 0.045

- **Reduce the rates for daytime minutes**
- Introduce more **flexible customisable pricing plans**

## International Plan

- Differentiate the international plan
- Gather feedback from current international plan users
- Develop tailored international plans



# NEXT STEPS

---

## Customer Service Calls

- Collect and analyse data on customer service calls:
  - reason for call
  - type of issue
  - was the issue resolved
  - customer service experience satisfaction

## State

- Collect and analyze additional data (feedback from customers) on reasons for churn in each state



# Further Recommendations

Introduce new features for analysis, including Internet data usage and Internet data services

---

Introduce new features on customers demographics (age, sex, family plan or not, number of people in family)

---

Include qualitative analysis for customer feedback to broaden understanding of features correlations



# THANK YOU

Emma Scotson: [github.com/emmascotson](https://github.com/emmascotson)

Dolgor Purbueva: [github.com/dolgorp](https://github.com/dolgorp)