By Samuel Stoltenberg

# Machine Learning for Customer Churn Prediction

Predicting why customers left the company

# Telecom Customer Churn Data

The data used was from a telecom company, with information like `phone number`, `day minutes used`, and length an account has been established.

The data was downloaded from kaggle under the title `Churn in Telecom dataset`

# Questions to Answer

1. What geographical locations should the company focus on?

2. Why does a customer decide to leave their current provider?

3. What type of model will perform best on the given dataset?

4. How can the company improve their customer retention?

By Samuel Stoltenberg

# OSEMN Process:

Obtain

Scrub

Explore

Model

Interpret

scikit
learn

# Models Used:

K Nearest Neighbors Classifier

Gradient Boosting Classifier

Random Forest Classifier

Support Vector Classifier

Adaboost Classifier

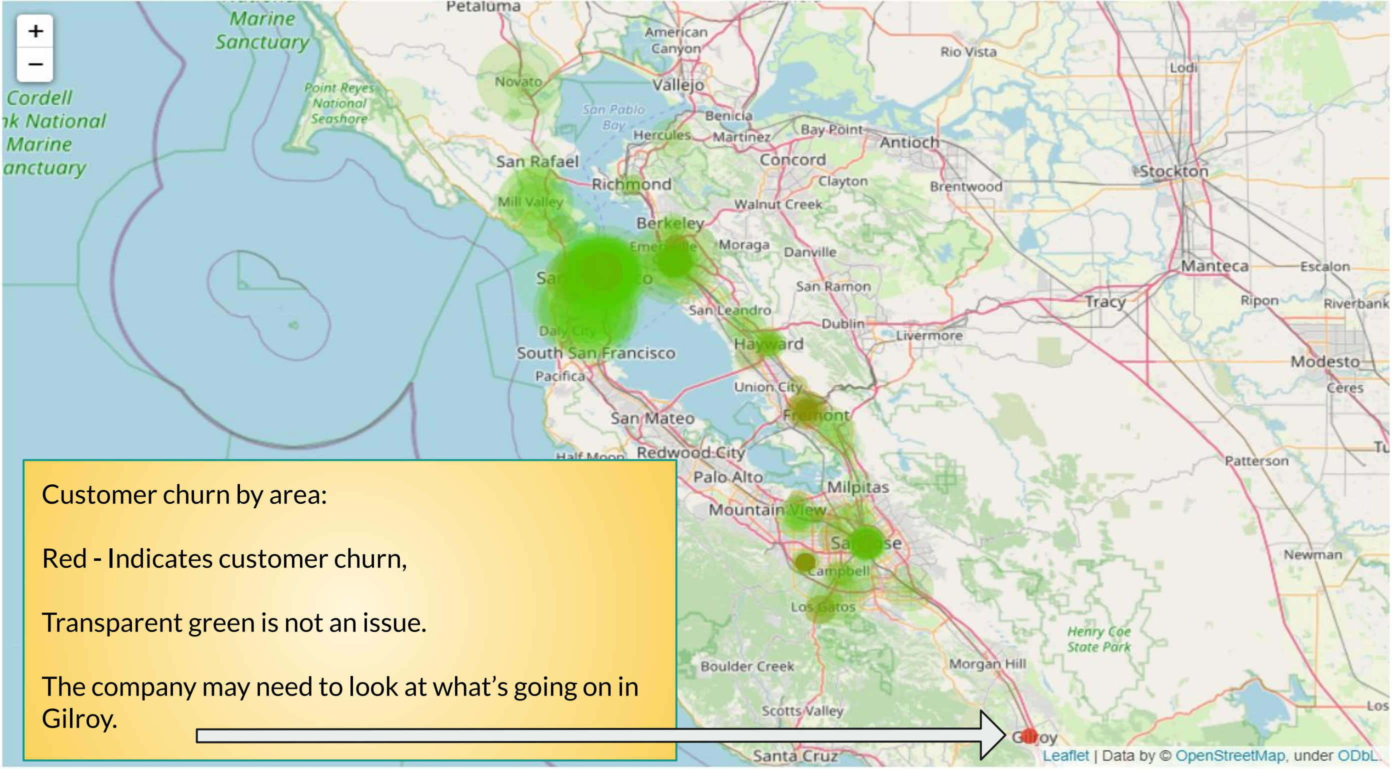
## Scrubbing the Data:

- There were no null values in the data.
- Some yes/nos that needed to be encoded into 1/0 respectively.
- Class imbalance needed to be dealt with, as only 15% of the data was on customers who had churned.

# Explore:

We used phone number for location scraping the for relative longitude and latitude of a customer for the map ahead.

Note: This could be skewed by someone moving, but keeping their phone number, as in the data we have multiple states, but all of the area codes are from California.



Customer churn by area:

Red - Indicates customer churn,

Transparent green is not an issue.

The company may need to look at what’s going on in Gilroy.

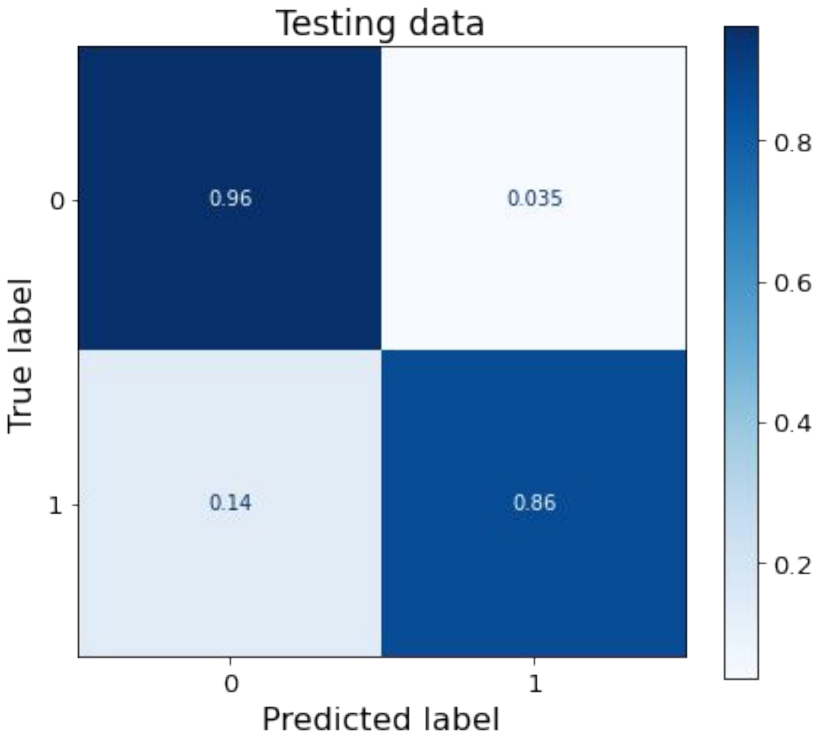
# Gradient Boosting Classifier:

After testing the models outlined in the fourth slide this is the model that performed best, giving us a overall testing score of ~96%,  and ~86% of the time being able to say whether a customer would leave or not.

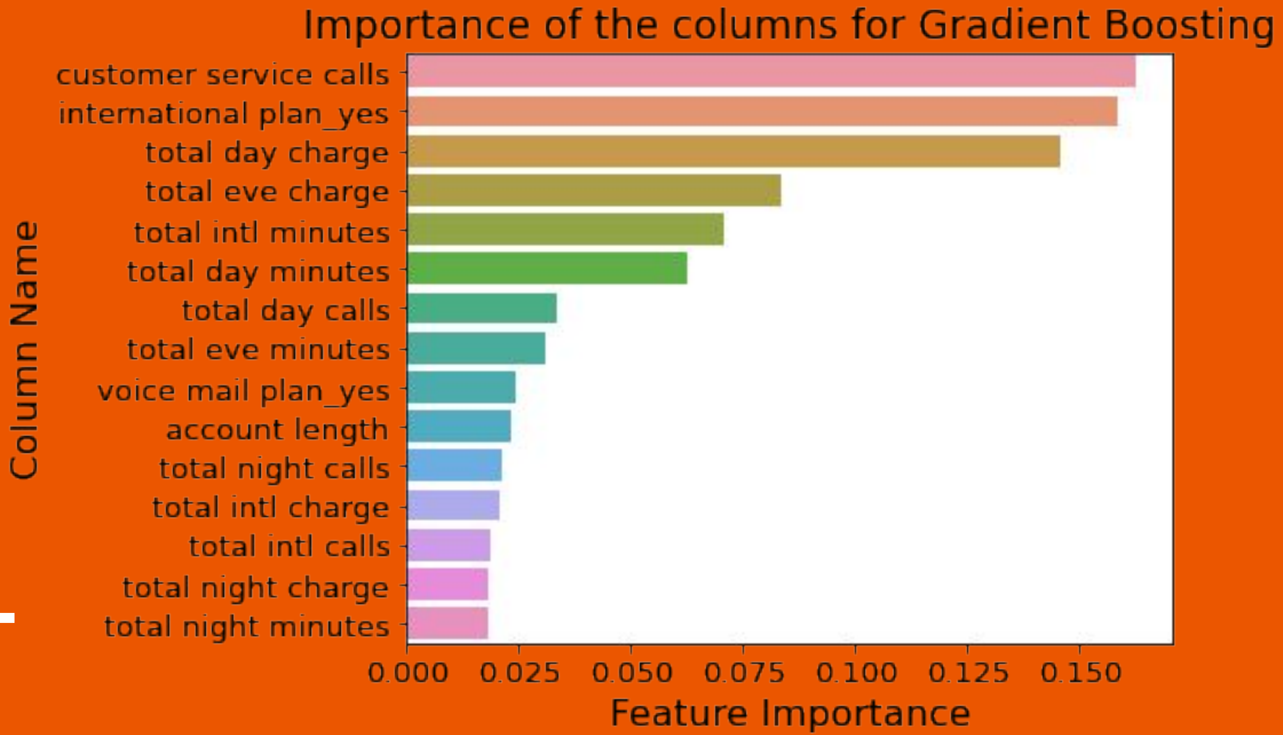
Don’t be confused by the squares on the right:

Bottom right is how many customers leaving our model successfully predicted.

Top left is how many customers staying our model successfully predicted.

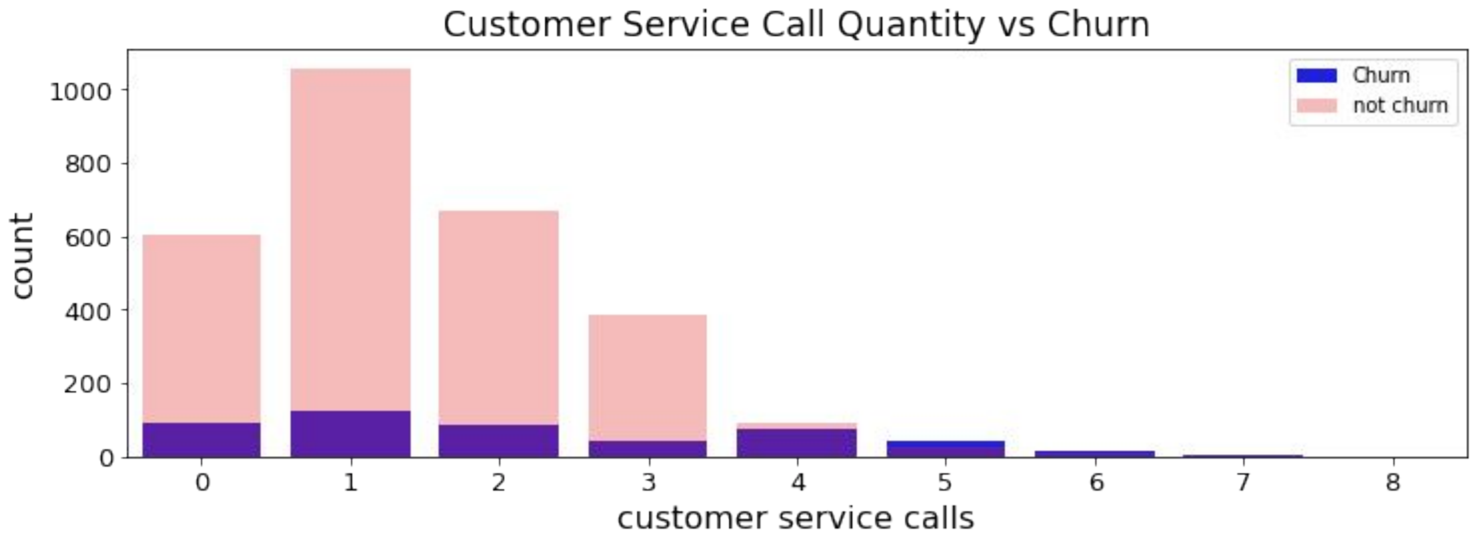


# Feature Importance:



# Customer Service Calls:

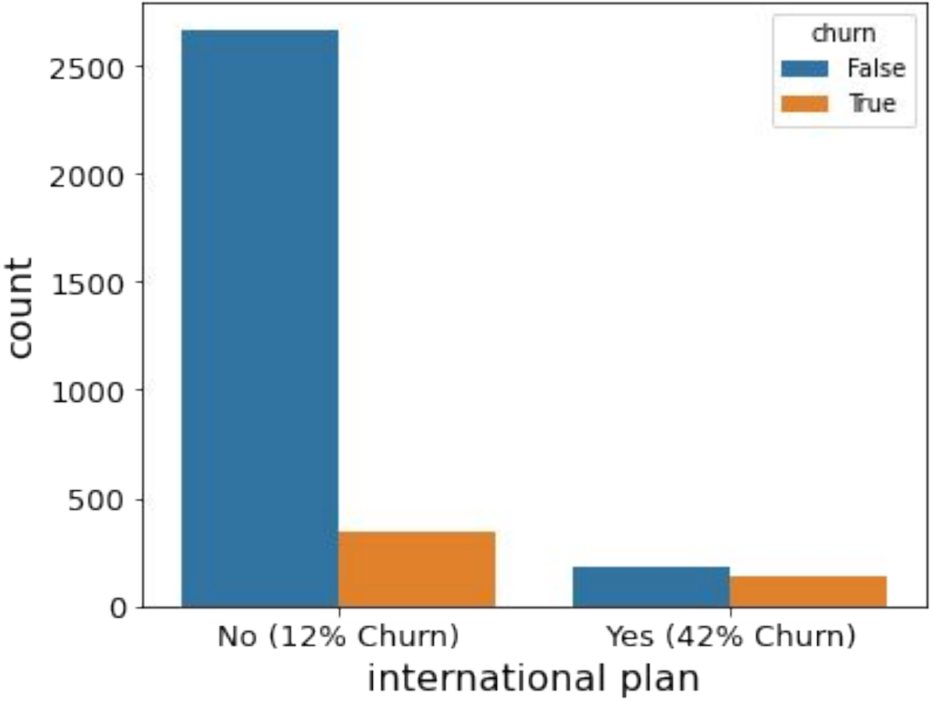
It seems that the more customer service calls a customer is making, the more likely they are to leave the company or “churn”.



# International Plan

The graph is showing that if someone has the international plan they are 30% more likely to churn than someone who does not.

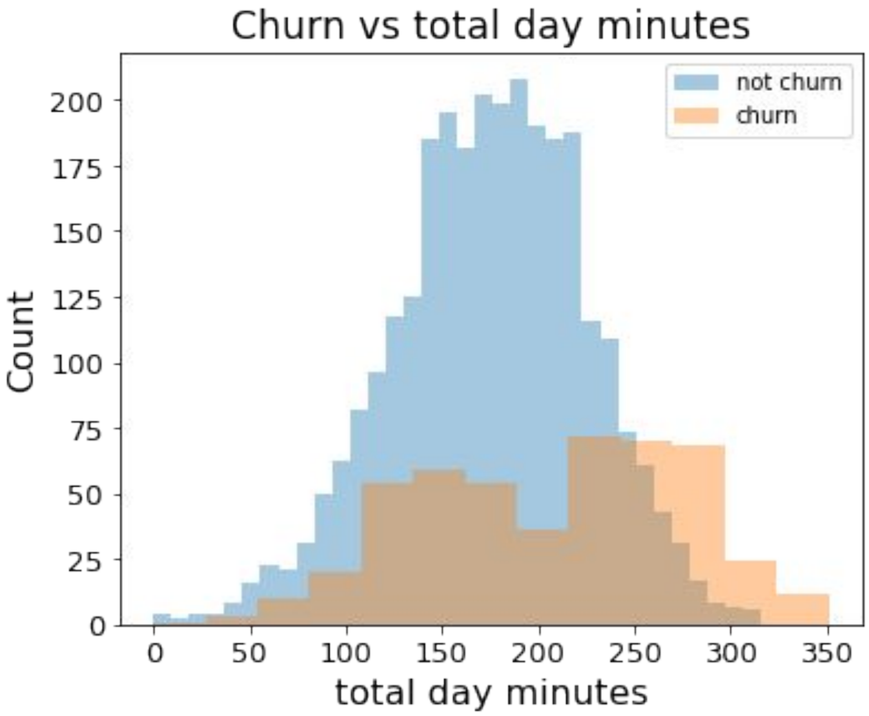
The company should lower their international plan rate, or find a different way to keep those customers.



# Total Day Minutes

The graph is showing that the more minutes a customer uses the more likely they are to leave the provider.

As total day minutes correlates to total day charge the company should think about lowering their rates.



# Conclusion:

- Out of the models we tried Gradient Boosting performed the best on the given data.
- The company should focus on their customer service calls, and how much they are charging for their minutes.
- We found that a customer is 30% more likely to churn if they have an international plan, and maybe the company should lower their international rate.
- The company should focus on the issues of customers calling in their 4th+ time, and see what the issues facing them is, as they are almost guaranteed to churn.

# Future Work:

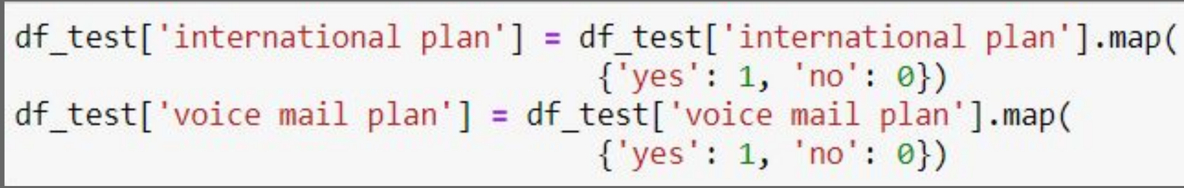
- Spending more processing power tuning the models, as all the models ran in under 30 seconds.
- Try different models like XGBoost, Extra Trees, and Logistic Classification.
- Running a Linear Regression on some of the important features to extract coefficients.

# Thank you.

# Appendix

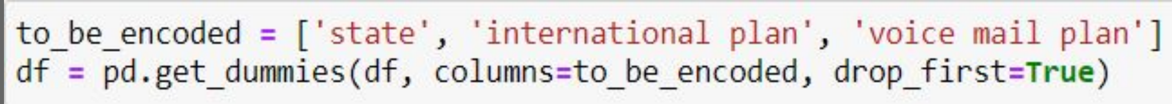
# Encoding:

| Sample [“international plan”] | | |
|---|---|---|
| yes | | 1 |
| no | | 0 |
| no | => | 0 |
| yes | | 1 |
| no | | 0 |



```
df_test['international plan'] = df_test['international plan'].map(
                                              {'yes': 1, 'no': 0})
df_test['voice mail plan'] = df_test['voice mail plan'].map(
                                              {'yes': 1, 'no': 0})
```

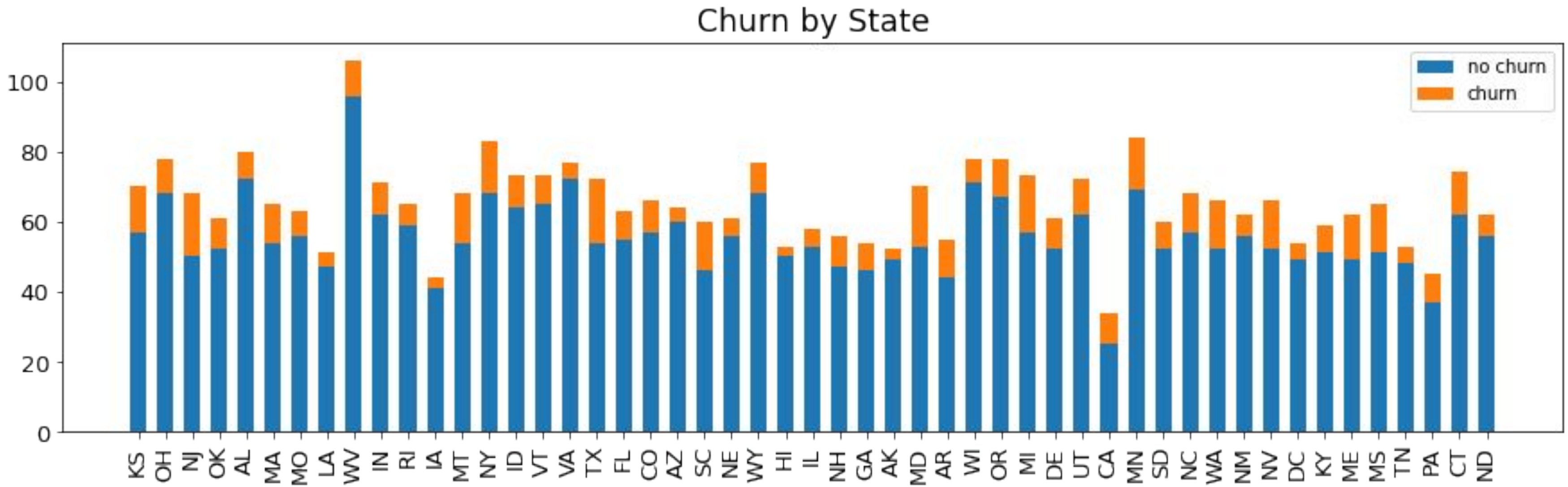
Pandas’ get_dummies function could also do this, but the column name would change from ‘international plan’ => ‘international plan_yes’



```
to_be_encoded = ['state', 'international plan', 'voice mail plan']
df = pd.get_dummies(df, columns=to_be_encoded, drop_first=True)
```

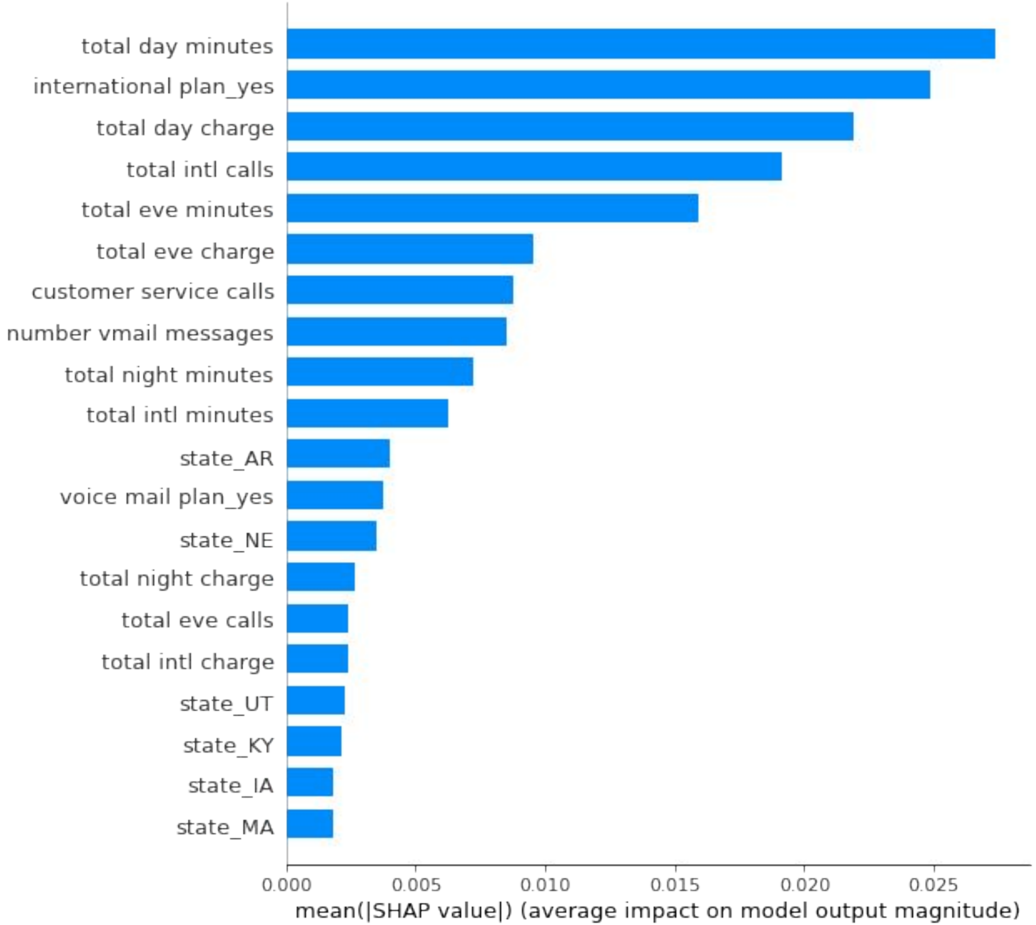
We do this because a 0 or a 1 is easier for a computer to understand vs a yes or a no.

# Churn Distribution:



# Shap feature importance

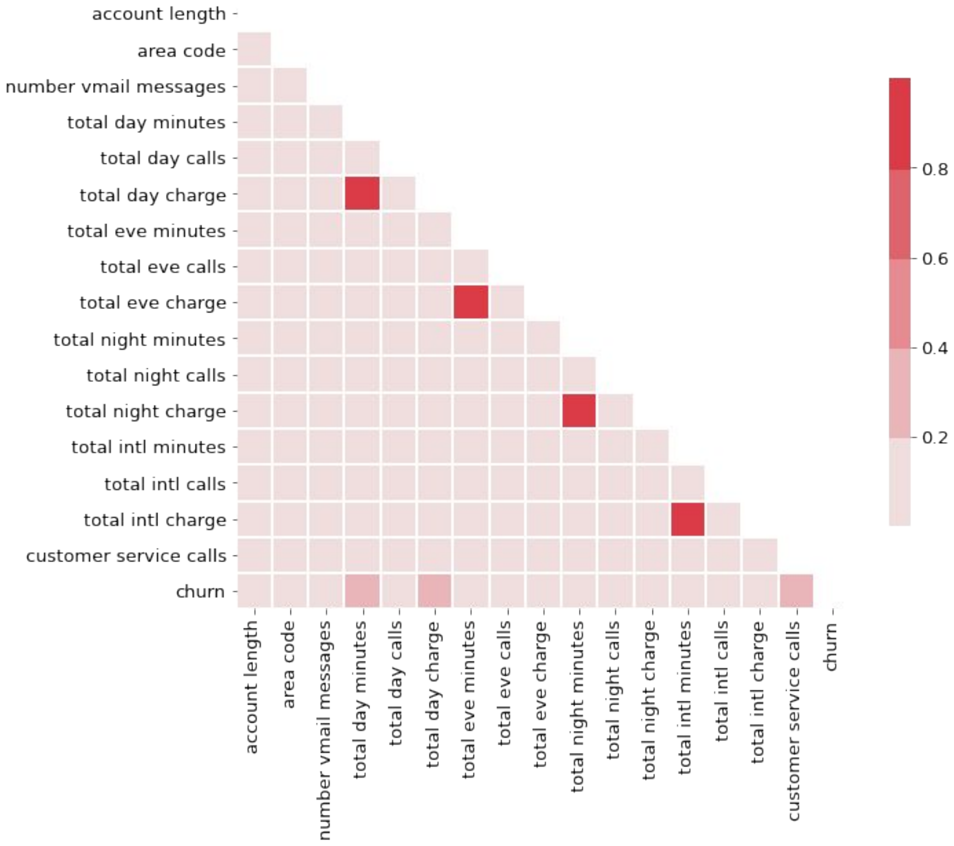
Just like the other importance chart day minutes and international plan are the most important.  A key difference from the other importance  is customer service calls is not at the top.



# Feature Correlation:

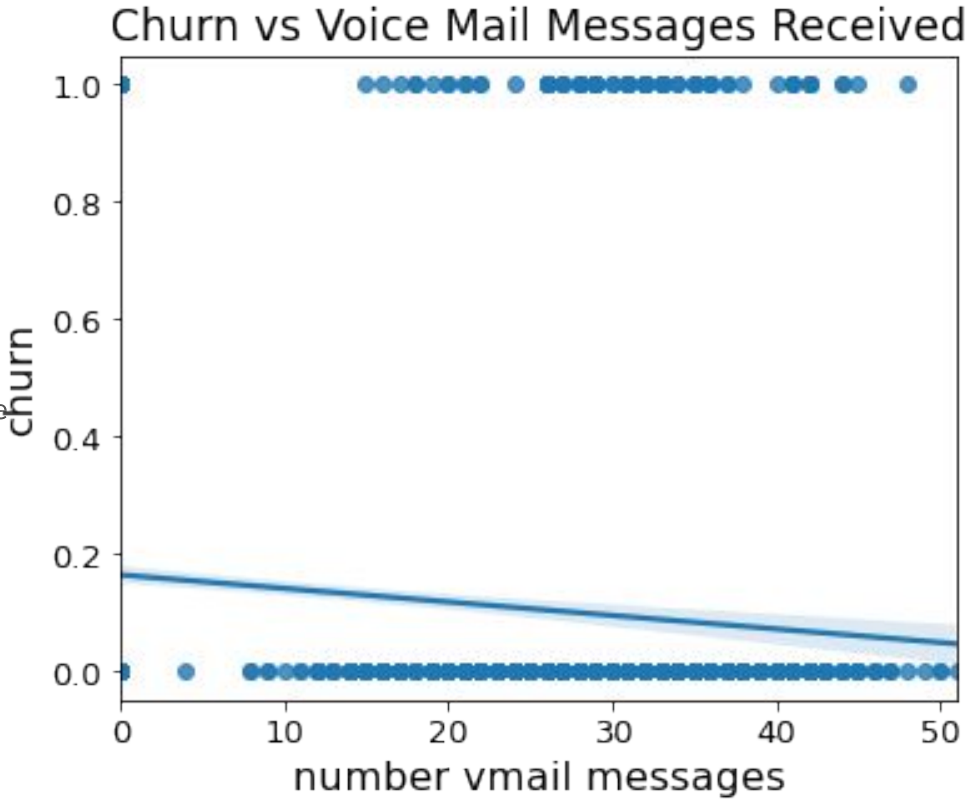
This is a pay per minute service as total minutes, and total charge are strongly correlated, which explains why customer churn goes up as minutes go up their bill is higher.

Total day minutes, charge, and customer service calls all slightly correlate with churn.  As shown in the feature importances.



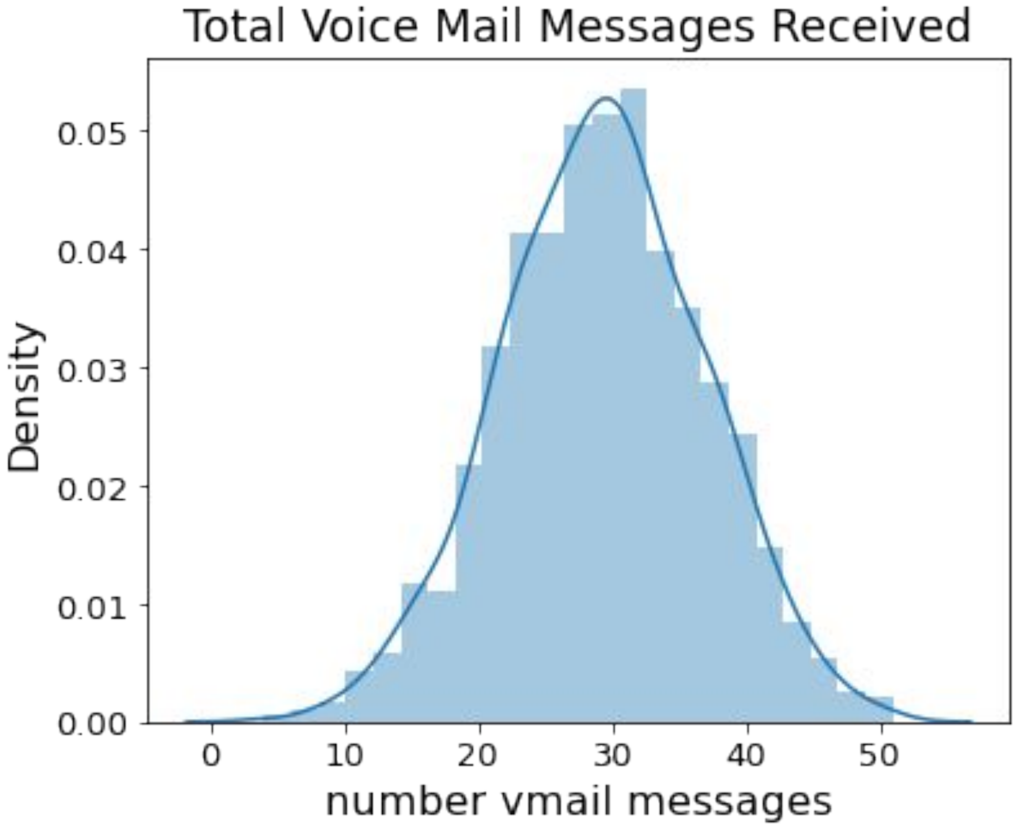
# Voice Mails

Maybe the company should leave more voice mails if they are not currently, as the more voice mails someone receives, the more likely they are to stay with the company.



# Voicemail Distribution:

On average, people receive more voicemails than I do.


Total Voice Mail Messages Received
Density
number vmail messages