Communication Tools

Prof. Philipp Borchert

Group:

Aazad Ghoslya
Sumanth Sripada
Ashwani Nitharwal

Data Description

We are provided with two dataset files churn_train, churn_test. The dataset contains multiple columns which tell us about the customers behavior and usage of various types of the plans. Based on the plans, customer use and the amount of usage of minutes, we need to predict the customer tendency to churn out of the company.

Process

We imported the dataset churn_train, churn_test as train and test.
For models to work well on the features and to remove the irrelevant features, we have performed some manipulations on the data. We dropped 2 columns 'state' and 'area_code'. We then encoded and replaced the 'churn', 'international_plan', 'voice_mail_plan' column values from 'yes', 'no' to 0, 1 for better functioning of models. We set the features as all the columns except 'churn' and assigned target column as 'churn' column. We then split 'churn_train' dataset into x_test and x_train. We further initiated the classes as

```
models = {
  "decision tree" : decision_tree,
  "logistic" : logistic,
  "neural_net" : neural_net,
  "random_forest": random_forest
}
```

and then fit the train as well as test data for the above models using the loop such as:

```
#fitting the models
for model in models:
  models[model].fit(X_train,y_train)
  print(f"{model} has been trained successfully")
```

We further predicted the above models and calculated the accuracy scores of the models. And have stored them in the dataframe to show as below

For splited train data:

	decision tree	logistic	neural_net	random_forest
Accuracy	1.0	0.85916	0.848739	1.0

Similarly, we performed predictions and calculated accuracy for the test data.
For splited test data:

	decision tree	logistic	neural_net	random_forest
Accuracy	0.929412	0.865098	0.851765	0.967059

Based on the accuracy of different models, we have selected to proceed with black box model which is random forest and one linear model which is logistic model with the good accuracy in prediction for this dataset. After deciding the model which is efficient, now we wanted to figure out which features are most influential ones over this model.

To interpret the two models i.e. random forest, logistic regression to find out the most contributing features for the customer churn prediction, we have used multiple interpreting techniques:

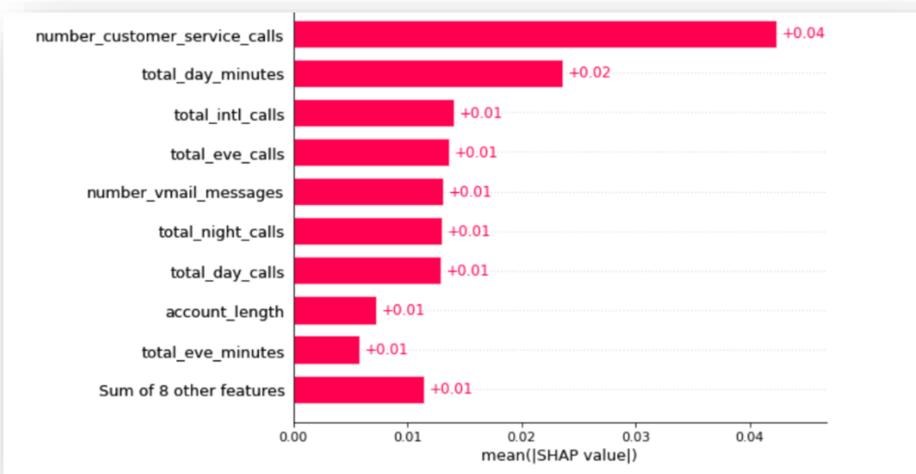
Shaply, Partial Dependence Plot(PDP) , Accumulated Local Effects(ALE) and Individual Conditional Expectation(ICE).

Interpreting the Logistic Model

We further performed calculations and plotted different types of SHAP value charts for all the features> Based on the below shap values, we have considered 6 most contributing features for the customer churn

Shaply for logistic model

bar chart



From the shap plot , we can see based on the below features, model has predicted the customer churn as per the weightage of the features

number_customer_service_calls

total_day_minutes

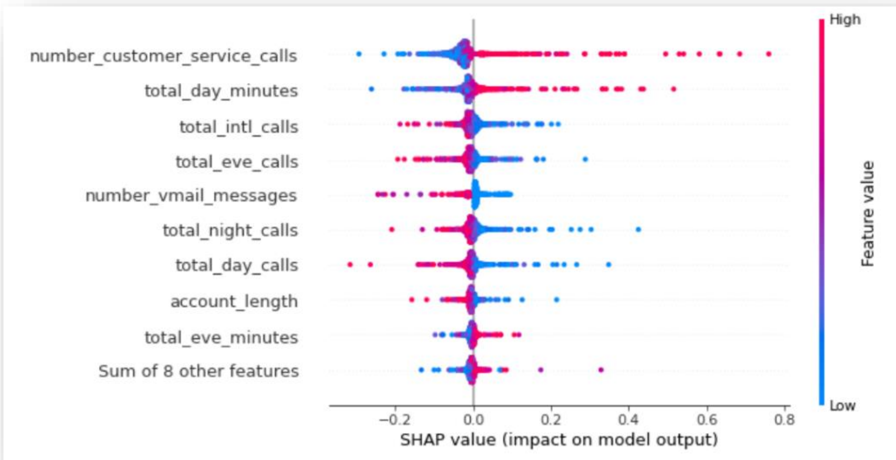
total_intl_calls

total_eve_charge

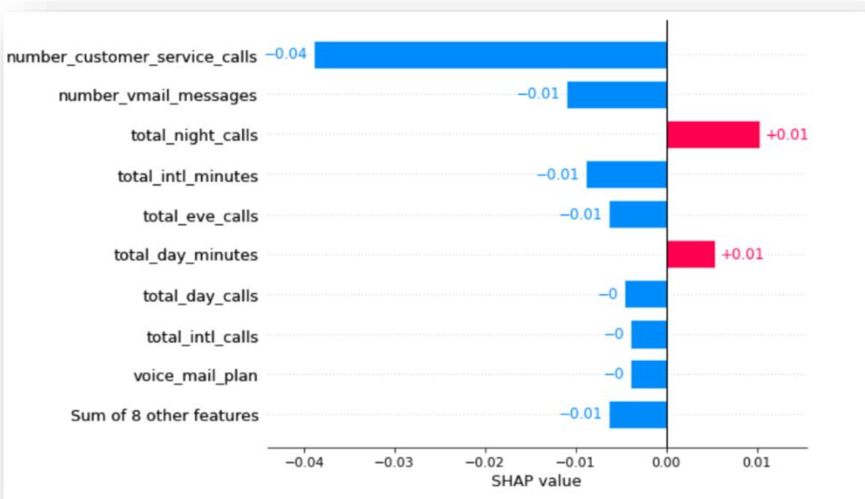
number_vmail_messages

total_night_calls

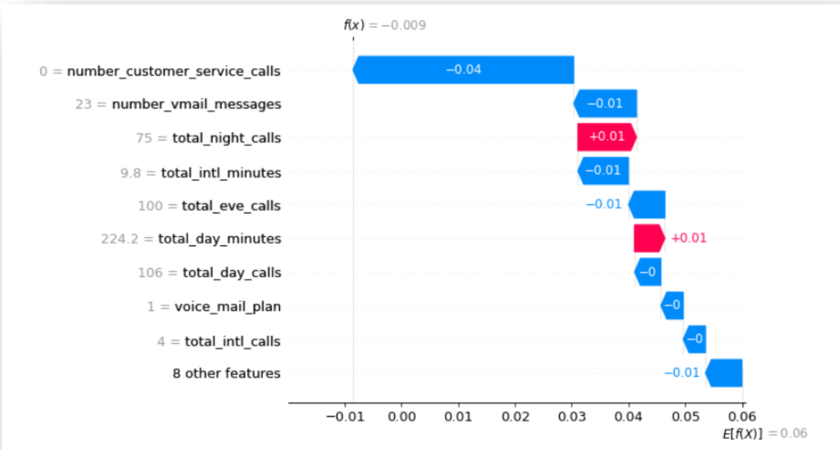
individual dots for each instance (beaswarm)



feature importance (cohorts)

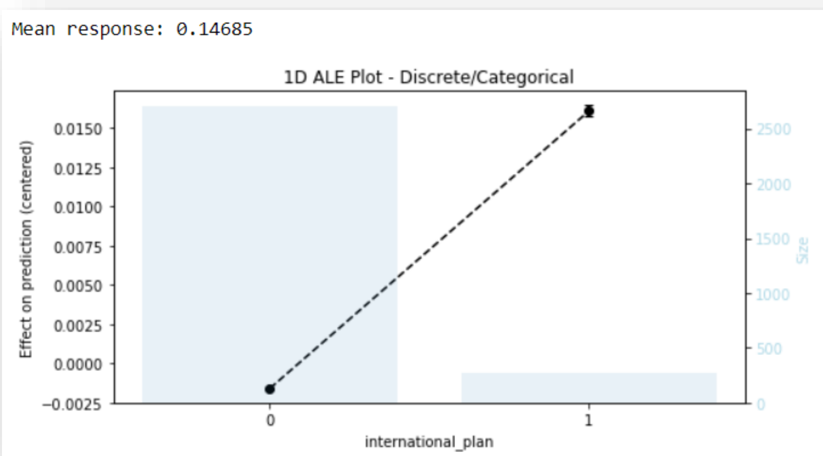


Waterfall model (feature weightage)

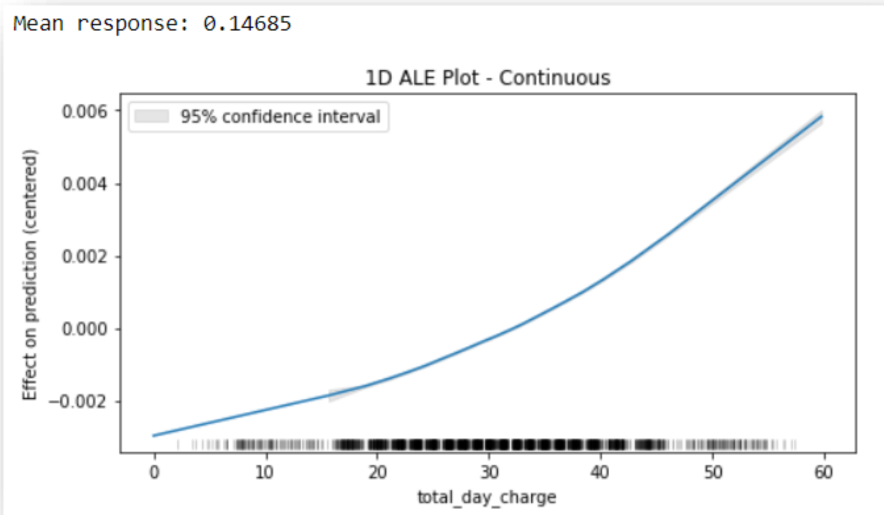


ACCUMULATED LOCAL EFFECTS Plot for logistic model

Below We can see the ACCUMULATED LOCAL EFFECTS plots for all the individual features based on the logistic model.

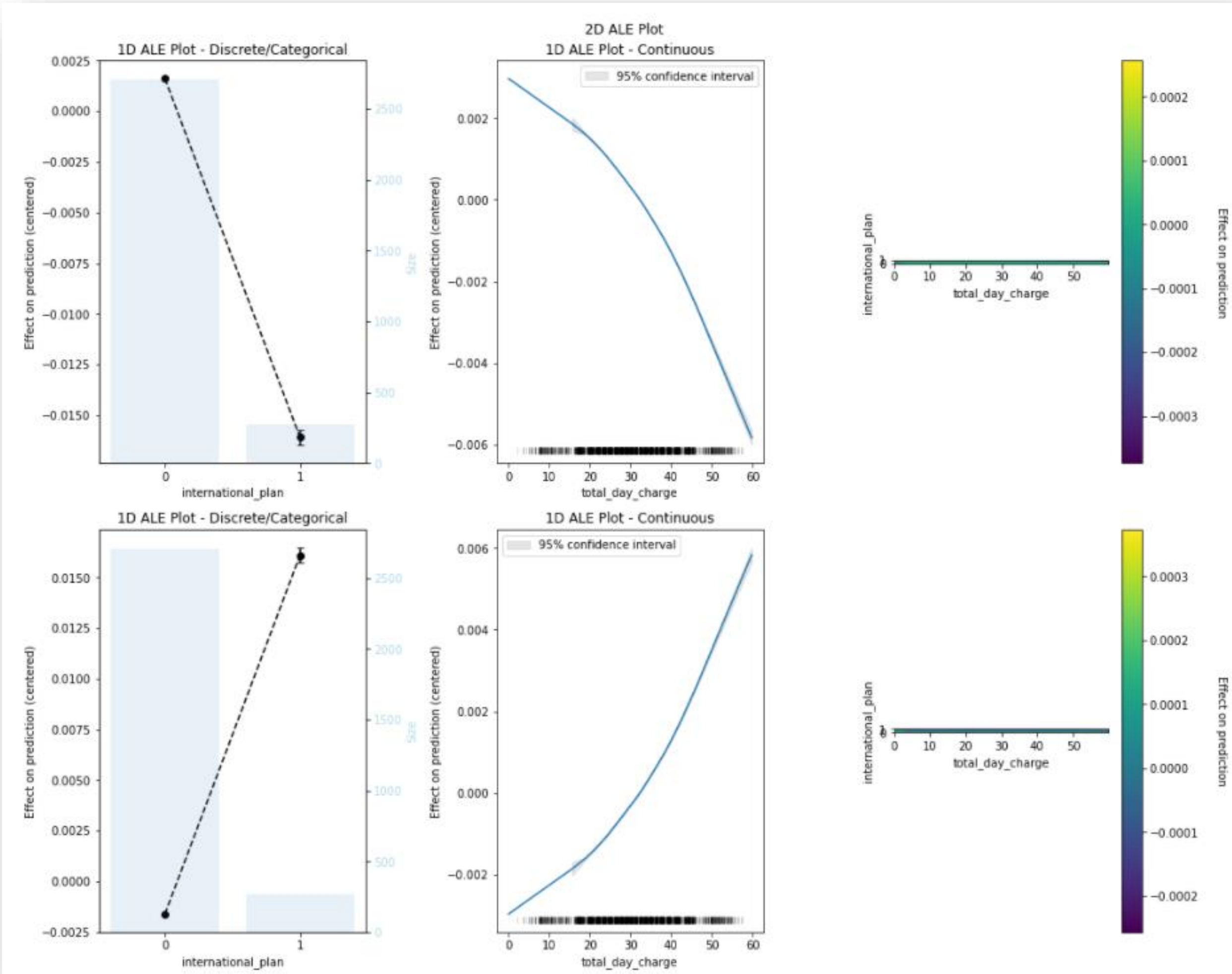


From this graph, we can see the number of customers having and not having international_plan and the effect on prediction by same



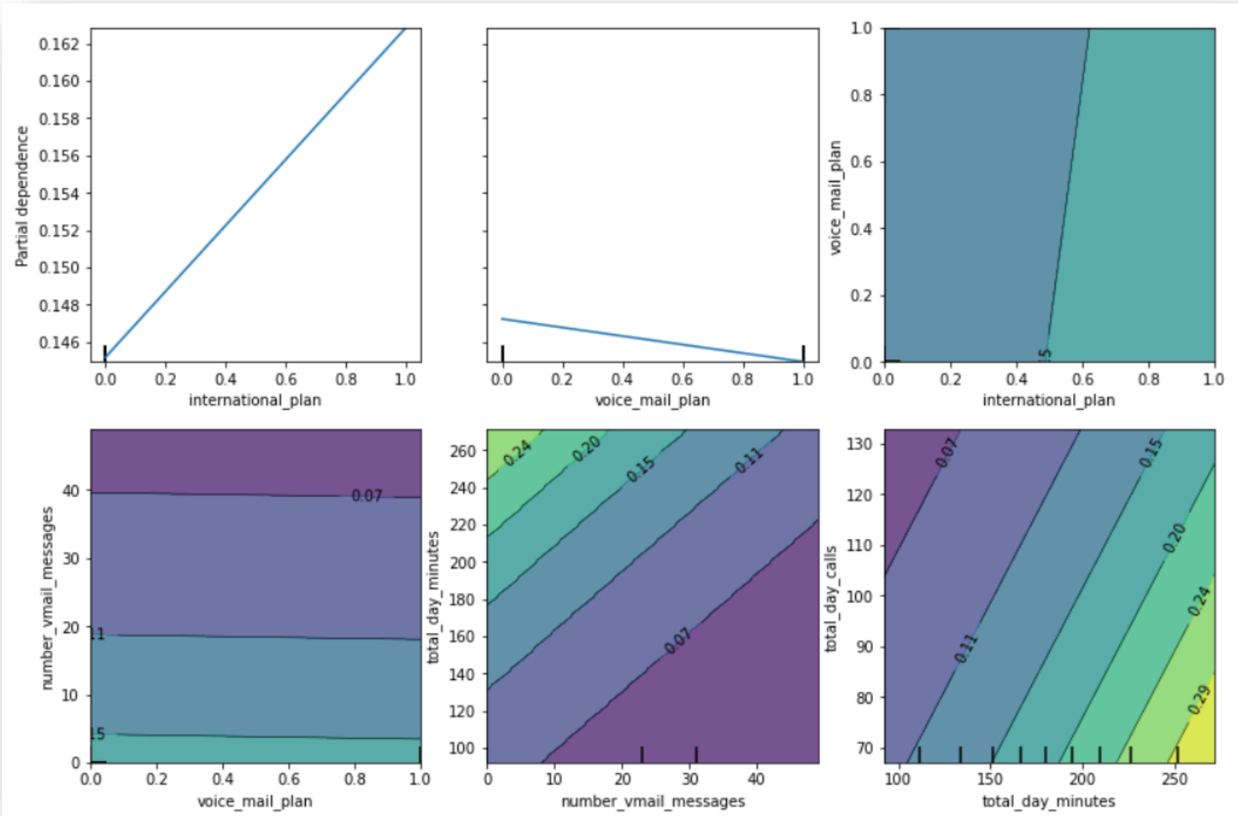
At 95% confidence, prediction is getting directly impacted by total_day_charge. The effect of prediction increases by increase in total_day_charge.

ACCUMULATED LOCAL EFFECTS plots for selected featuresThettt



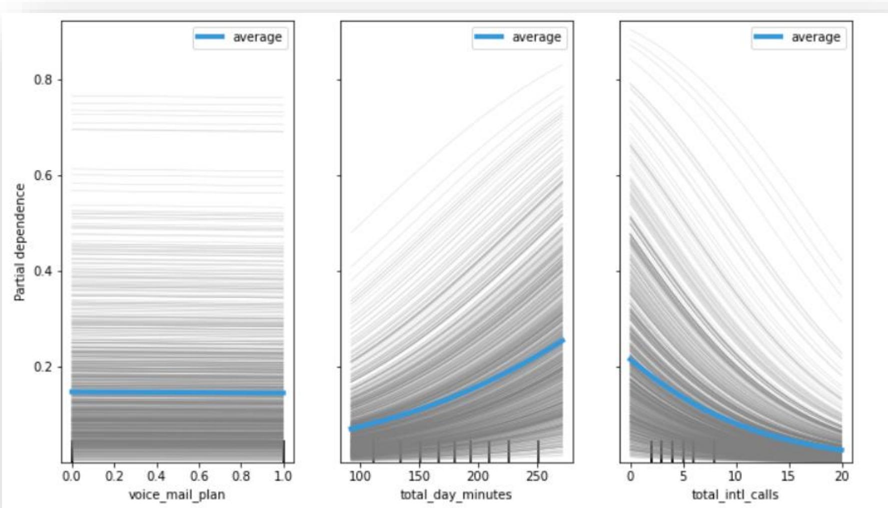
From these plots, we can see these mentioned features have direct impact on customer churn prediction.

PARTIAL DEPENDENCE PLOT for logistic model



From these plots, we can see these mentioned features have direct impact on customer churn prediction.

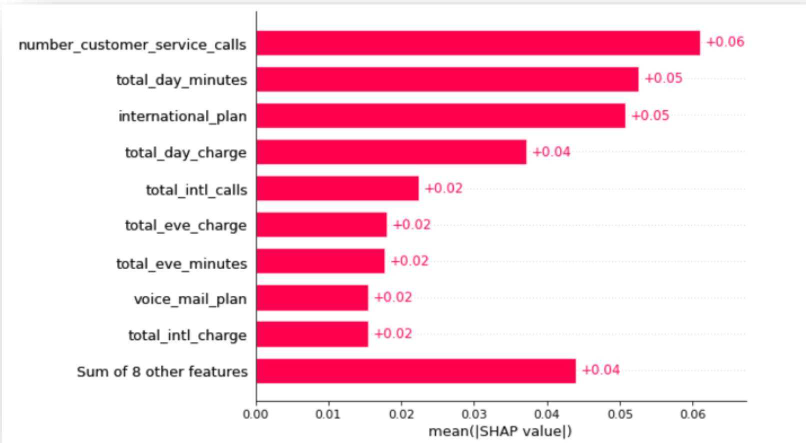
INDIVIDUAL CONDITIONAL EXPECTATION for logistic model



From charts, we can see the dependency of different features and their means that show the direction of impact.

Random Forest Model

Shaply for random forest

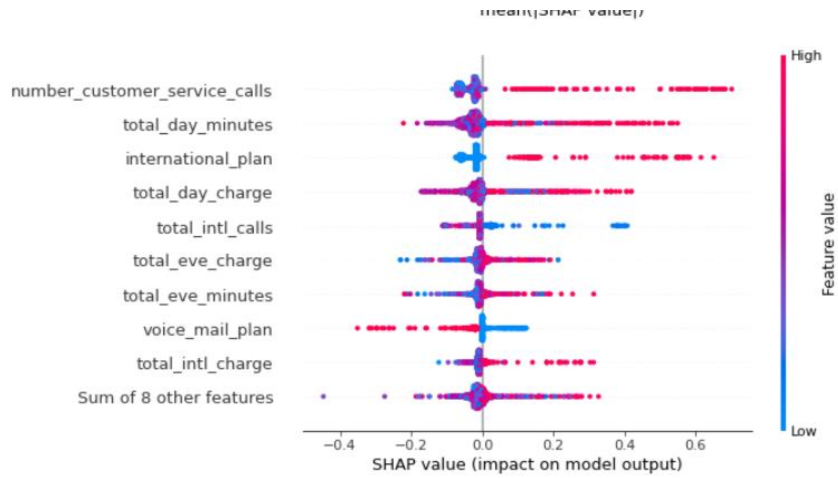


From the Shap values of Random Forest Model we can see the below features are the most contributing for the customer churn.

number_customer_servIndividual Conditional Expectation_calls
total_day_minutes
total_intl_calls
total_eve_charge
international_plan
total_day_charge

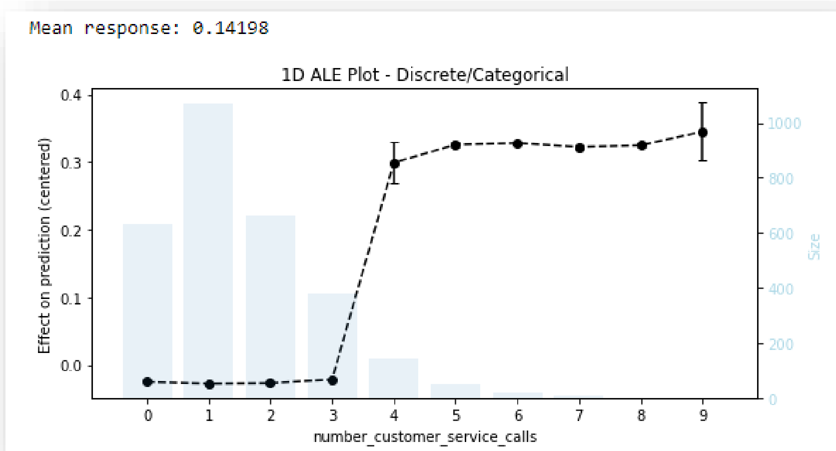
After interpreting both logistic and random forest models, and based on the shapely values we have selected the features which are common in both the models in contributing for the churn.

#individual dots for each instance

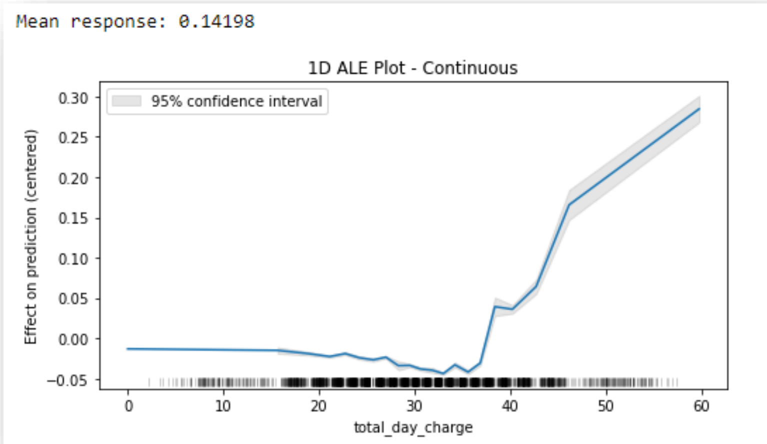


ACCUMULATED LOCAL EFFECTS for Random Forest

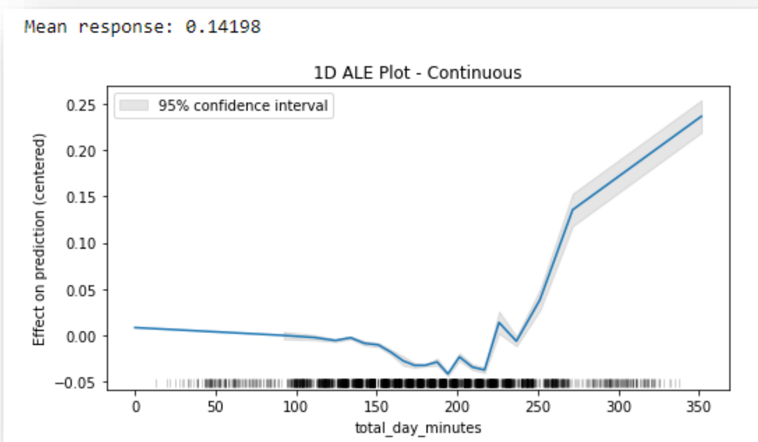
- ACCUMULATED LOCAL EFFECTS plots for all the selected individual features



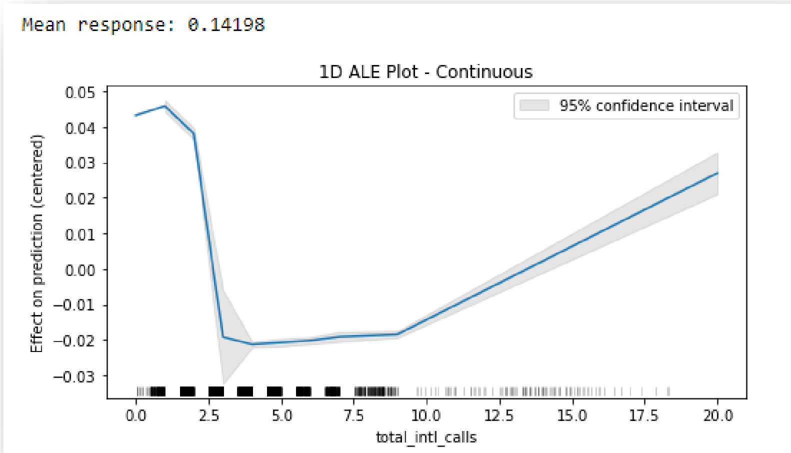
From these plots, we can see number_customer_service_calls feature have direct impact on customer churn prediction.



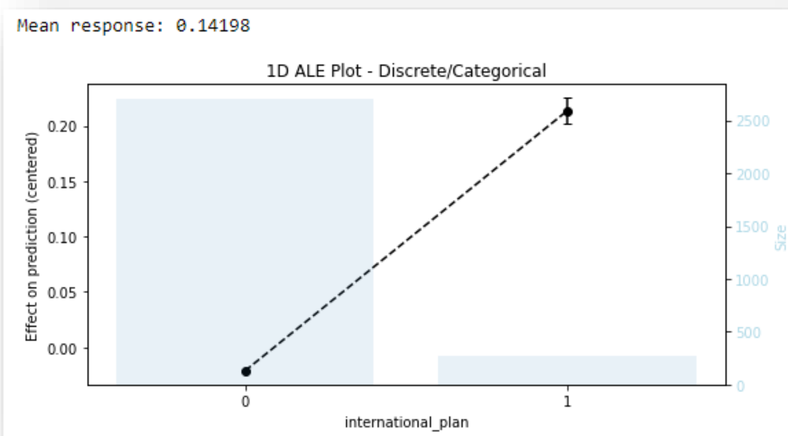
From these plots, we can see total_day_charge feature have direct impact on customer churn prediction.



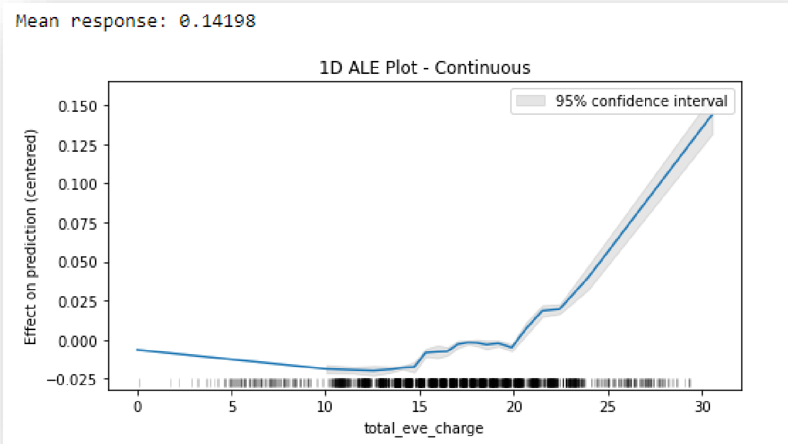
From these plots, we can see total_day_minutes feature have direct impact on customer churn prediction.



From these plots, we can see the total_intl_calls feature have direct impact on customer churn prediction.

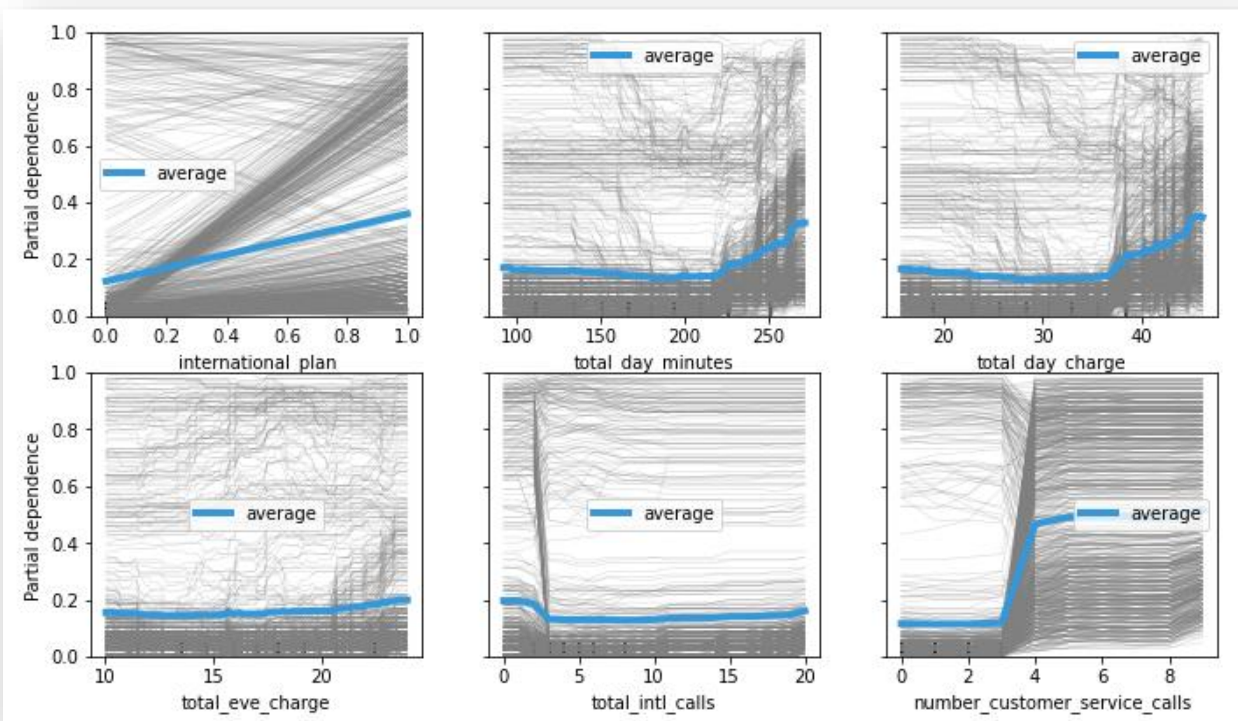


From these plots, we can see the international_plan feature have direct impact on customer churn prediction.



From these plots, we can see the total_eve_change feature have direct impact on customer churn prediction.

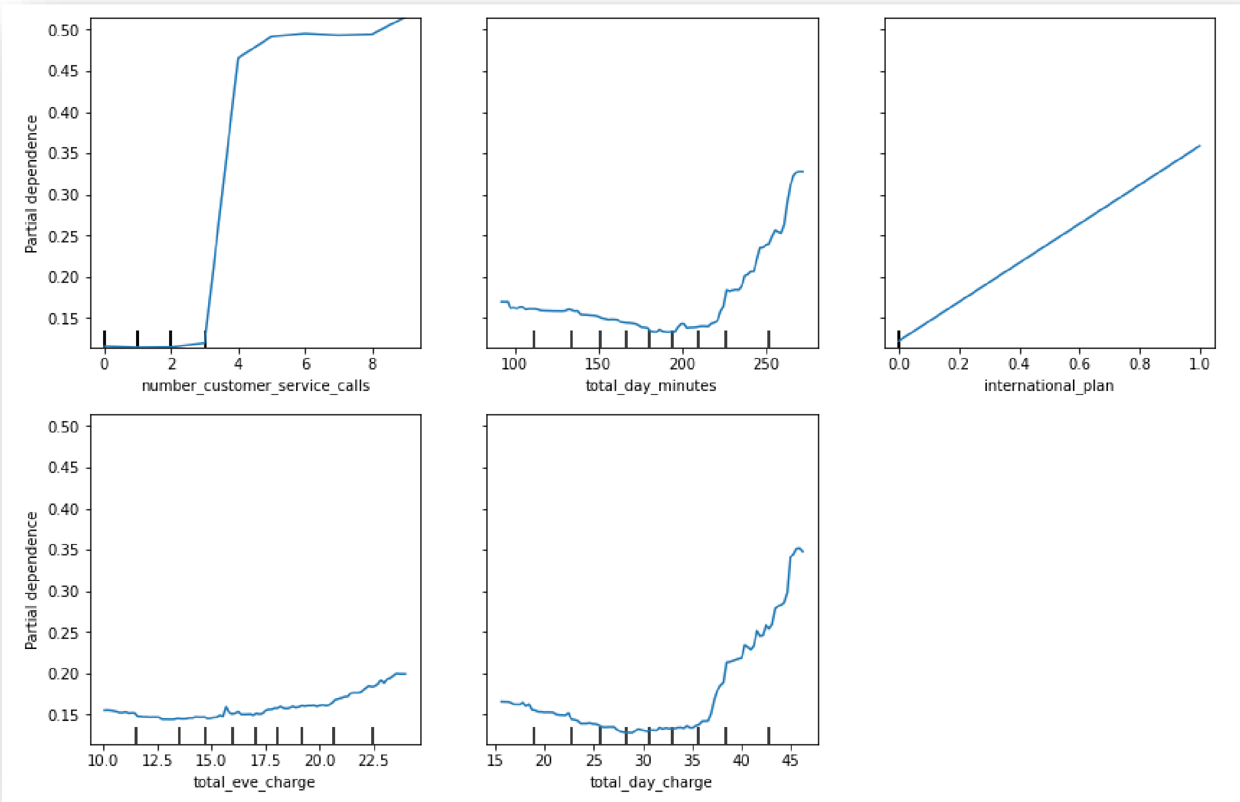
Interpreting Random Forest using INDIVIDUAL CONDITIONAL EXPECTATION



From these plots, we can see these mentioned features have direct impact on customer churn prediction.

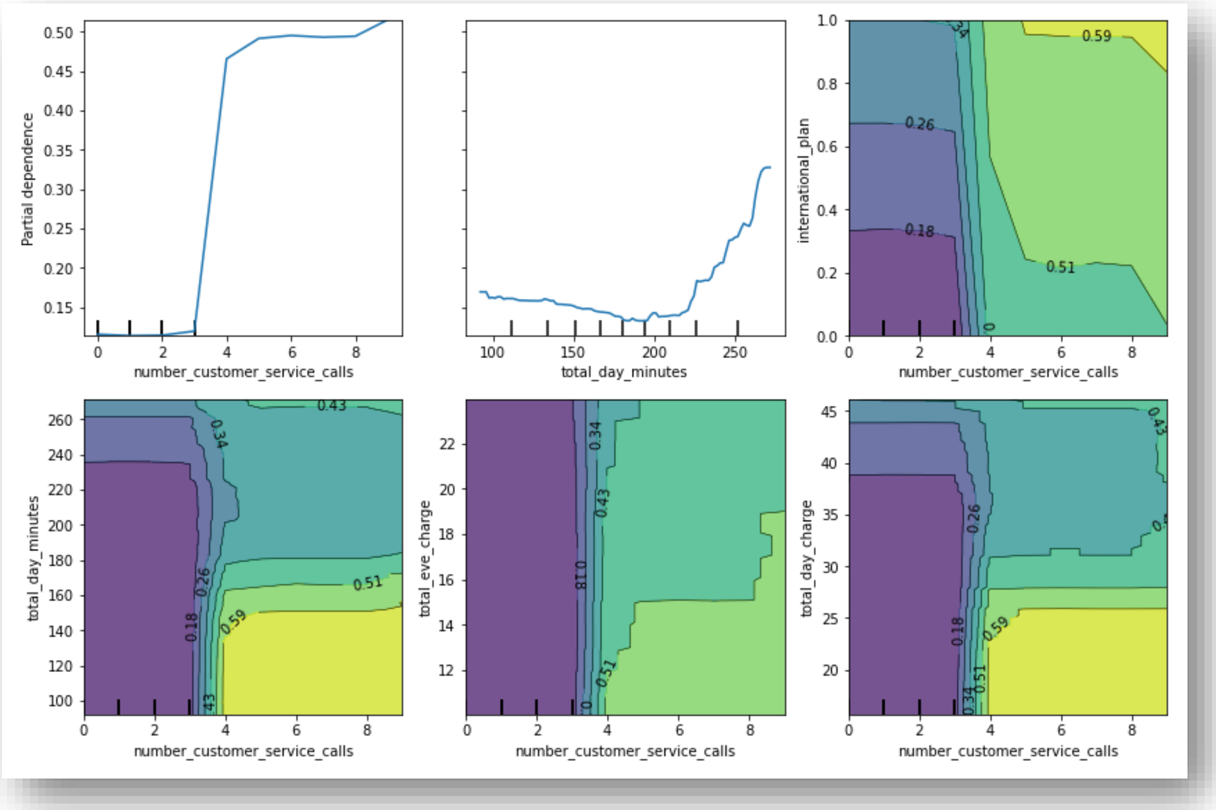
Interpreting Random Forest using PARTIAL DEPENDENCE PLOT

1D PARTIAL DEPENDENCE PLOT



From these plots, we can see these mentioned features have direct impact on customer churn prediction.

2D PARTIAL DEPENDENCE PLOT



From these plots, we can see these mentioned features have direct impact on customer churn prediction.

Based on the performance of both the modes we have tested the Random Forest Performance on the test dataset to see how the model predicts the customer churn on just the selected 6 features and not all features

We finally add these generated probabilities to the test table in a new column called 'churn'

```
test["churn"] = predictions
```

The final dataset is added with a column having predicted values as
0 - 'No Churn'
1 - 'Churn'
The final dataset looks like:

	number_customer_service_calls	total_day_minutes	total_day_charge	total_intl_calls	international_plan	total_eve_charge	churn
0	1	265.1	45.07	3	0	16.78	1
1	0	223.4	37.98	6	1	18.75	0
2	4	120.7	20.52	6	0	26.11	1
3	3	190.7	32.42	3	0	18.55	0
4	3	124.3	21.13	5	0	23.55	0
...	...	...	...	...	...	...	...
745	0	119.4	20.30	7	0	19.24	0
746	3	177.2	30.12	2	0	22.99	0

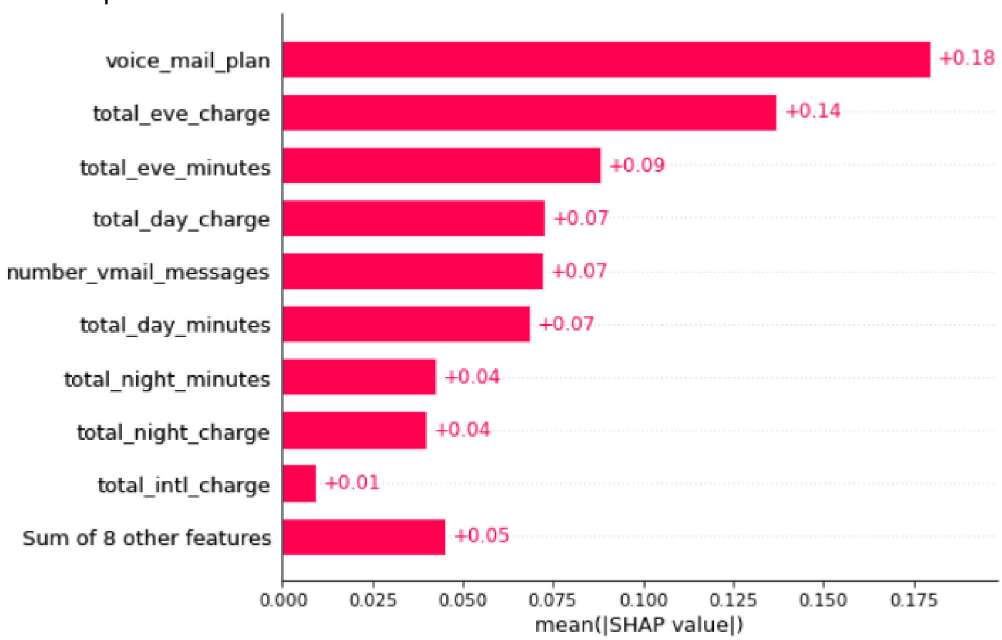
- To better understand the performance of our Model, We have taken the conditional subset of the data set where total day minutes is greater than 250 mins.

Below is the model performance on the conditional subset data:

	logistic	random_forest
Accuracy	0.795276	0.92126

For the Subset data, we have interpreted the model again to understand the behavior of the model and have observed the minimal difference in the performance.

- Shap Values for the subset data



- For All the respective codes and the plots , please refer to the attached Jupyter notebook along with the dashboard.