Assignment 3 Report

Telecom Churn

- **Outlook Login details:**

  Username : adssummer2016team1@outlook.com

  Password : team1team1

- **Website link :** http://adsteam1.azurewebsites.net/webform1.aspx
- **Github link :** https://github.com/abhijeet-s/ADS_Assignment/tree/master/Assignment3

~Team 1

1. Introduction & Problem Description

Business in the consumer market has to deal with churn time and again; Sometime churn is excessive and influences policy decisions. The traditional solution is to predict high-probability churners and address their needs via marketing campaign or providing special deals. These offers vary from industry to industry.

The problem at hand is to use historic telecom customer data and predict which customers have high probability of churning such that the organization can take measures to retain the customer by changing their marketing strategy or providing special deals.

2. Data Description

The dataset consists of 5000 rows and 21 columns.

The columns present in the dataset are:

- State – It is a discrete variable. This consists of the 50 states in US.
- Account Length – It is a continuous variable. This is the number of months a customer has been with the provider.
- Phone number - It is a discrete variable. This is the phone number of the customer. In this case this column should have unique values.
- International Plan - It is a discrete variable. This column tells us if a customer has opted for an international plan. The values can either be YES or NO.

- Voice mail Plan - It is a discrete variable. This column tells us if a customer has opted for an international plan. The values can either be YES or NO.
- Number vmail messages - It is a continuous variable. This tells us the number of voice mail messages the person received in a month.
- Total day minutes - It is a continuous variable. This tells us the total number of minutes a customer has made calls for in the daytime.
- Total day calls - It is a continuous variable. This tells us the total number of calls a customer has made in the daytime.
- Total day charge - It is a continuous variable. This tells us the total bill generated due to the day calls made.
- Total eve minutes - It is a continuous variable. This tells us the total number of minutes a customer has made calls for in the evenings.
- Total eve calls - It is a continuous variable. This tells us the total number of calls a customer has made in the evenings.
- Total eve charge - It is a continuous variable. This tells us the total bill generated due to the evening calls made.
- Total night minutes - It is a continuous variable. This tells us the total number of minutes a customer has made calls for at night.
- Total night calls - It is a continuous variable. This tells us the total number of calls a customer has made at night.
- Total night charge - It is a continuous variable. This tells us the total bill generated due to the night calls made.

- Total intl minutes - It is a continuous variable. This tells us the total number of minutes a customer has made international calls for.
- Total intl calls - It is a continuous variable. This tells us the total number of international calls a customer has made.
- Total intl charge - It is a continuous variable. This tells us the total bill generated due to the international calls made.
- Number customer service calls: It is a continuous variable. This tells how many calls the customer has made to customer support.

3. Data Preprocessing

Normalization of the variables:

Normalization of data is important to ensure that the variables measured at different scales contribute equally to the analysis. Owing to greater numeric range, the impact on response variables by feature having greater numeric range could be more that the one having less numeric range and this could, in turn, impact prediction accuracy. And the churn problem values predictive accuracy and not allow a particular feature impact the prediction due to large numeric value range. Thus, we normalized the dataset.

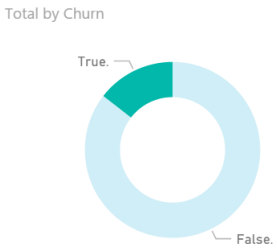
We used Min-Max normalization technique, which uses the formula:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

4. Exploratory Analysis

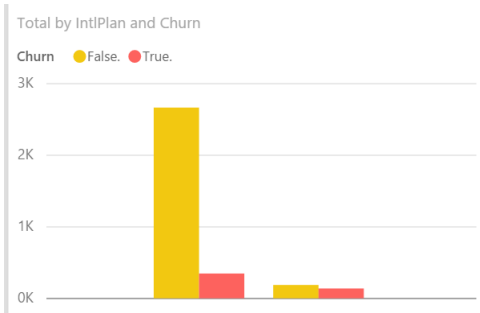
The exploratory analysis was performed on Power BI.

The current prevalence of churn is shown below:



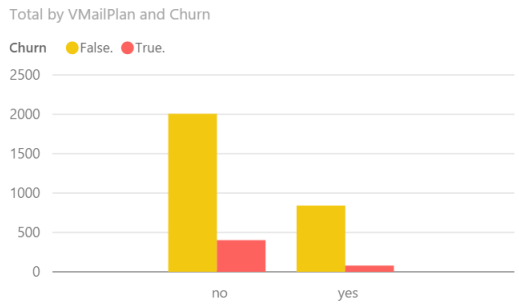

```
##
## False  True
##  0.86  0.14
```

So for every 100 customers in the sample, 14 opt to switch to other providers and 86 tend to continue with the same company, which is a significant number and thereby we need to determine the factors influencing customer churn rate as it costs much more to gain a new customer than to retain an existing one

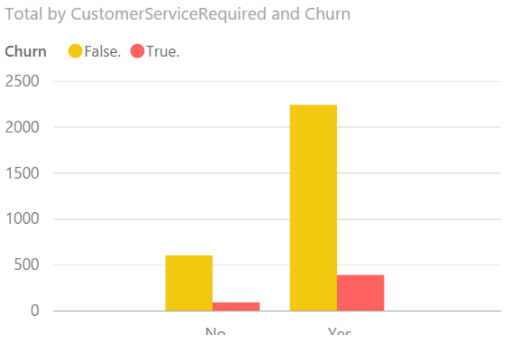


Of the people without any international call plans, ~10% of the customers leave, while considering those with international plan, almost 50%. So it's highly probable that the international plan might not be attractive for customers who require long overseas phone calls. Or the service of international calls might be poor, forcing customers to leave.

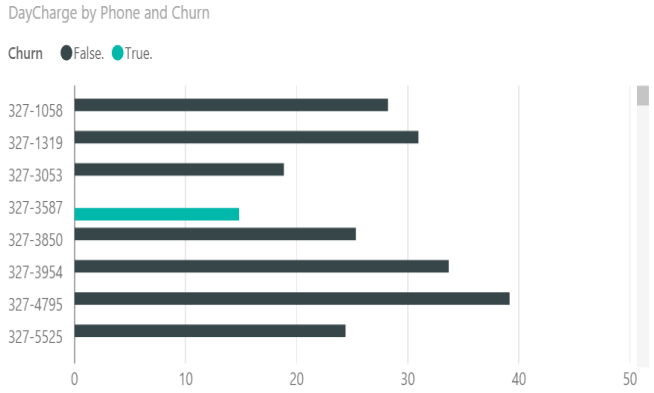
Now, upon quick look at the voice mail plan:

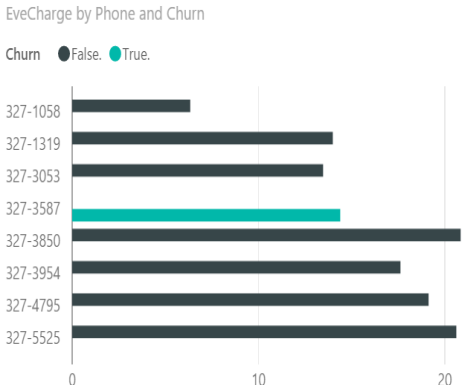


This shows that the voice mail plan seems to be very popular among the customers, with only 7% of the customers with the said plan leave, while close to 17% of the customers without voice plan leave. This could suggest that the company provides a reasonably good voice mail service, but for some reason not being opted by large proportion of customers. May be the customers are unaware of the voice mail plans or the plan costs high, not attracting customers. Hence, the company should probably focus on spending time and money in marketing Voicemail plans which can generate more revenue.



Not much can be inferred from this plot, although we can say that customer service quality is reasonably good.



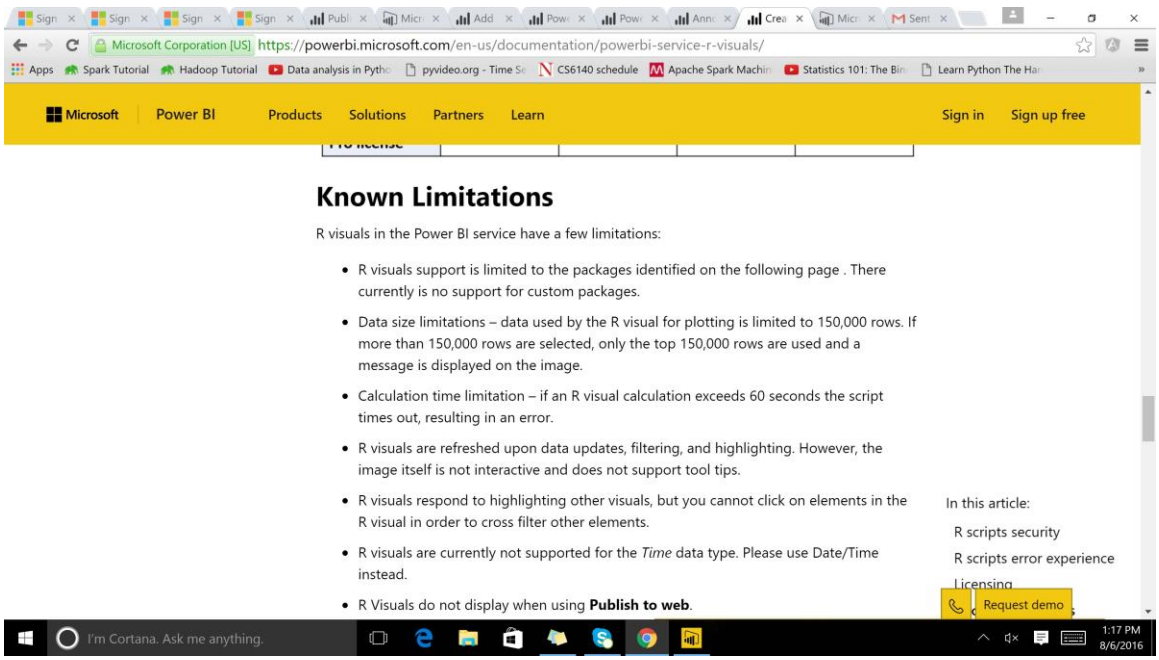


The day call charges are higher than that of evening and night. This might illustrate why the day time total call duration is an important factor for churn. It may well be that the customers get high charges and are not willing to put up with it. It could indicate that despite having lower total call durations that evening and night, day time call service seem to be adversely contributing to churn.

Link for Power BI :

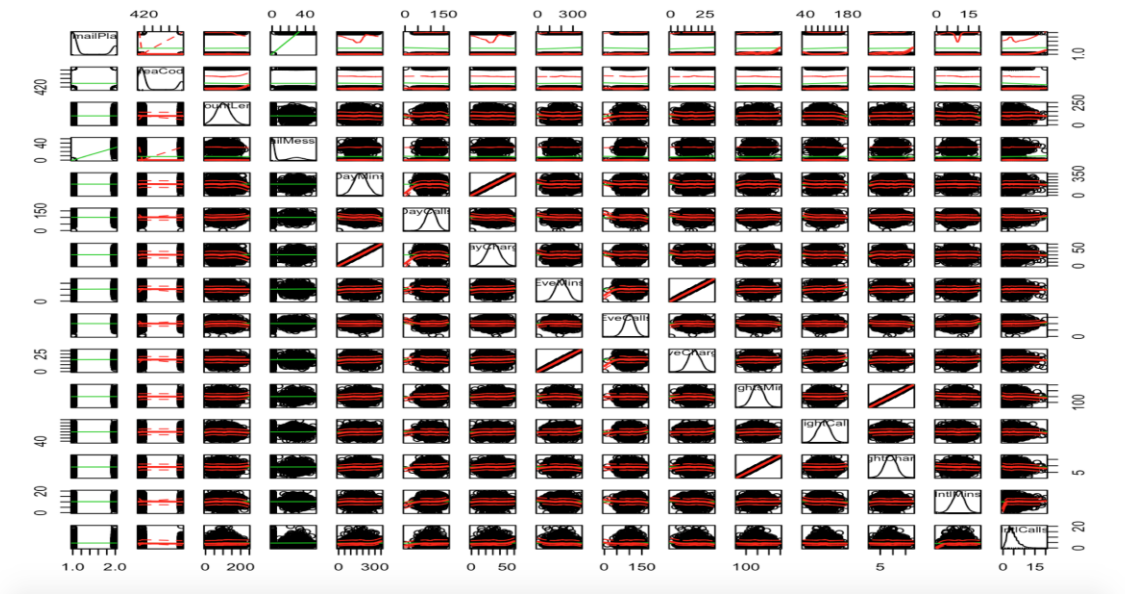
https://app.powerbi.com/view?r=eyJrIjoiNjI4NWNiZjMtYTEzZi00Nzk5LThlMTMtNTUxODY1YWM3MTI1IiwidCI6IjZhYmZjNzNmLWRhNjQtNDEzNy05ZjlmLTE1ZmFhZTU2ZjY4NSIsImMiOjN9

We have integrated R with Power BI and R visuals do not display when using Publish to Web.



Correlation between Variables:

Most of the variables in the dataset are uncorrelated. Four significant correlations were observed.

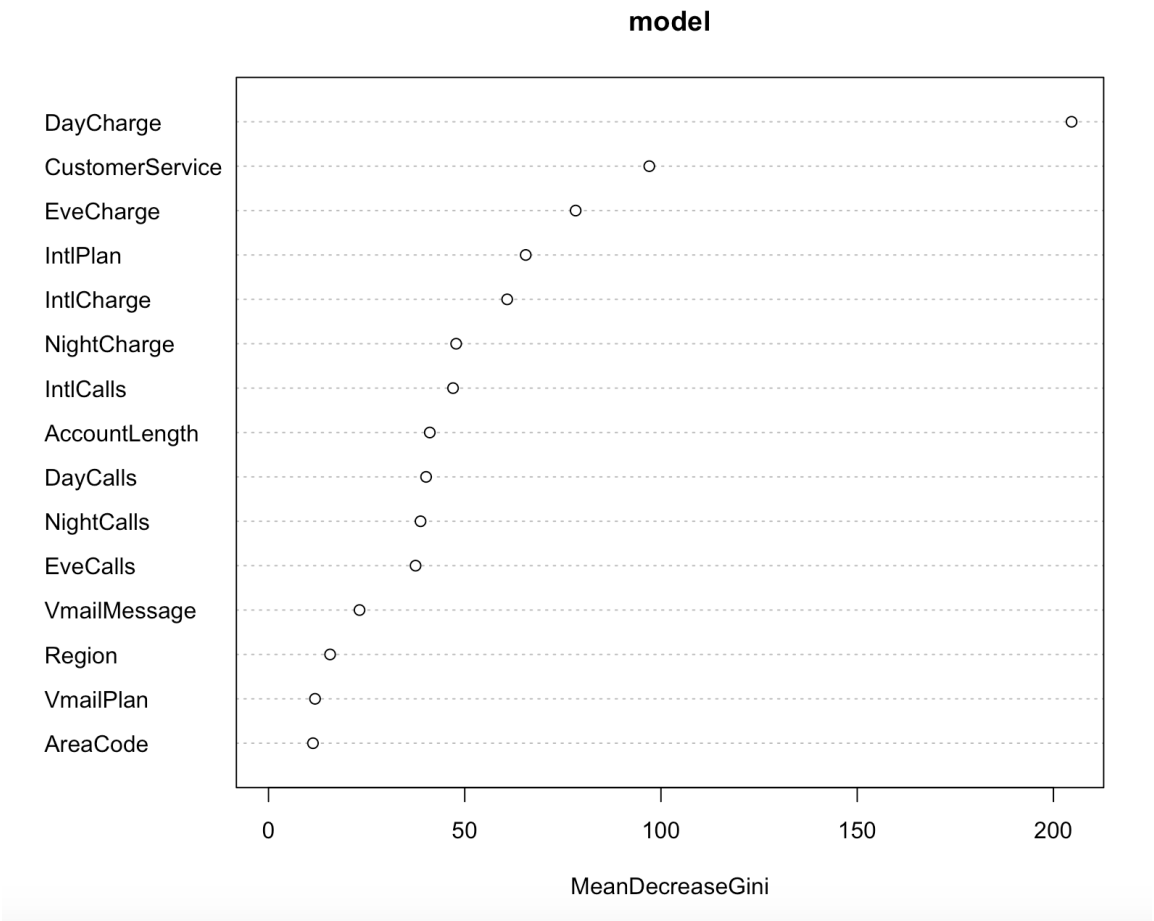


1. There is high correlation between Day Mins and Day Charge
2. There is high correlation between Evening Mins and Evening Charge
3. There is high correlation between Night Mins and Night Charge
4. There is high correlation between International Mins and International Charge

We removed the Day Mins, Evening Mins, and International Mins from our model, as it doesn't add any additional information.

Feature Selection:

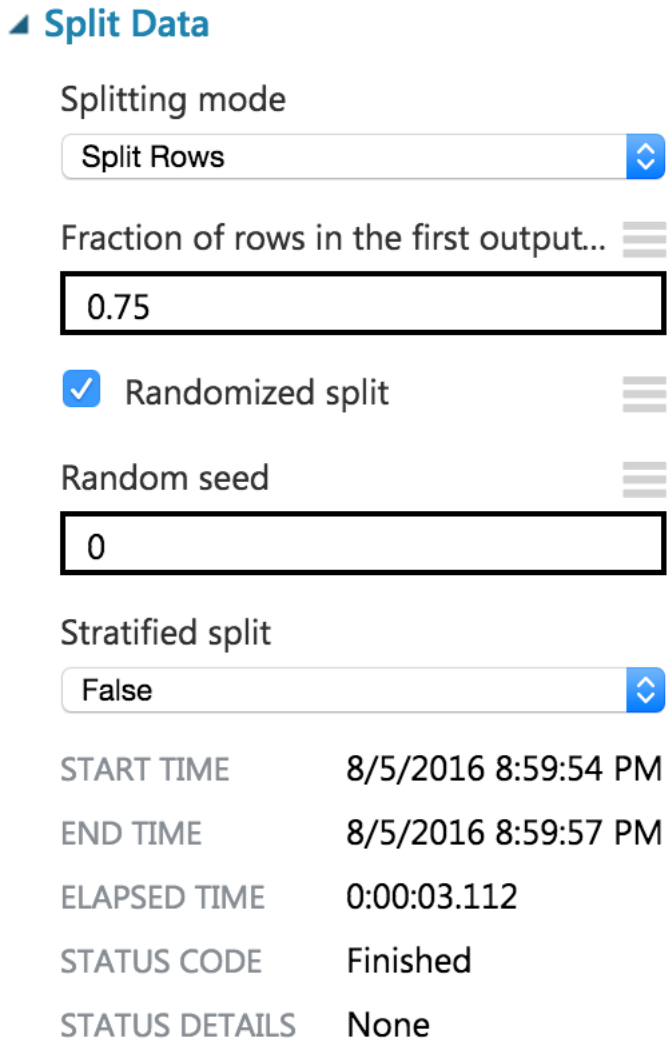
1. Initially looking at the data the column AreaCode was removed as there is a column State that provides the same data.
2. The column phone number was removed, as it doesn't hold any importance to predict if the customer will churn or not.
3. We added a new feature called region which would bucket the different states to regions namely – West, South, Midwest, Northeast
4. We used the RandomForest model to plot the Importance of the Features, which would predict churn.



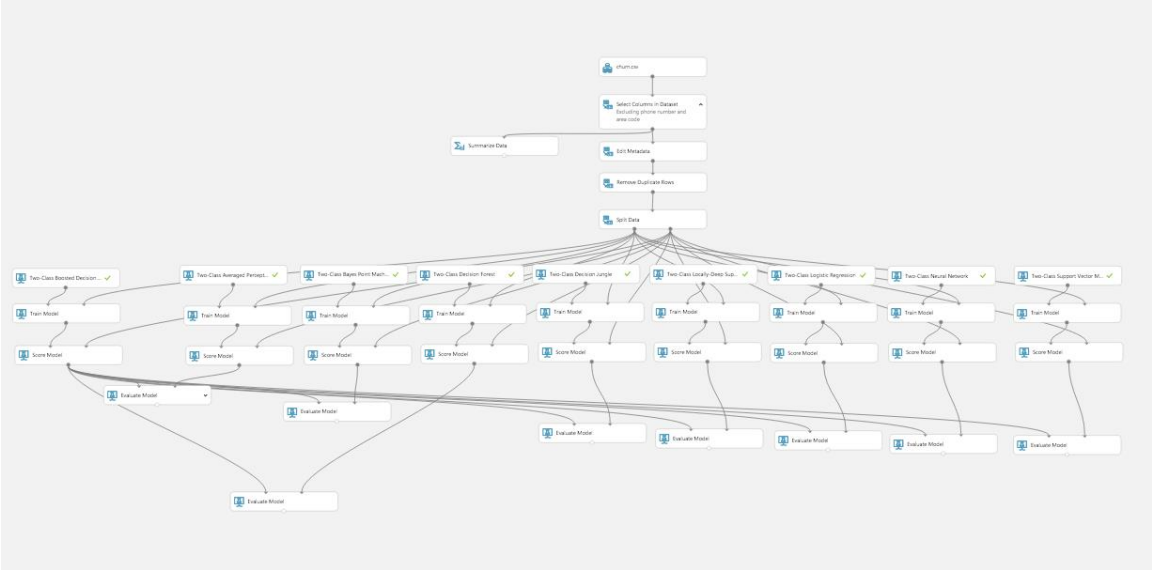
5. In the above graph you can see that Day Calls, Night Calls, Evening Calls have equal importance therefore added a new feature called Regular Calls and removed the Day Calls, Night Calls, Evening Calls.

5. Prediction Model

I. The data is split into Test and Train



II. We started by comparing every classification model available in Microsoft Azure



Model Name	R Square
Two Class Boosted Decision Tree	92.7%
Two Class Averaged Percepton	84.2%
Two Class Bayers Point Machine	84.9%
Two Class Decision Forest	93.4%
Two Class decision Jungle	93.4%
Two Class Locally- Deep SVM	90.7%
Two Class logistic Regression	85.2%
Two Class Neural Network	93.1%
Two Class SVM	83.2%

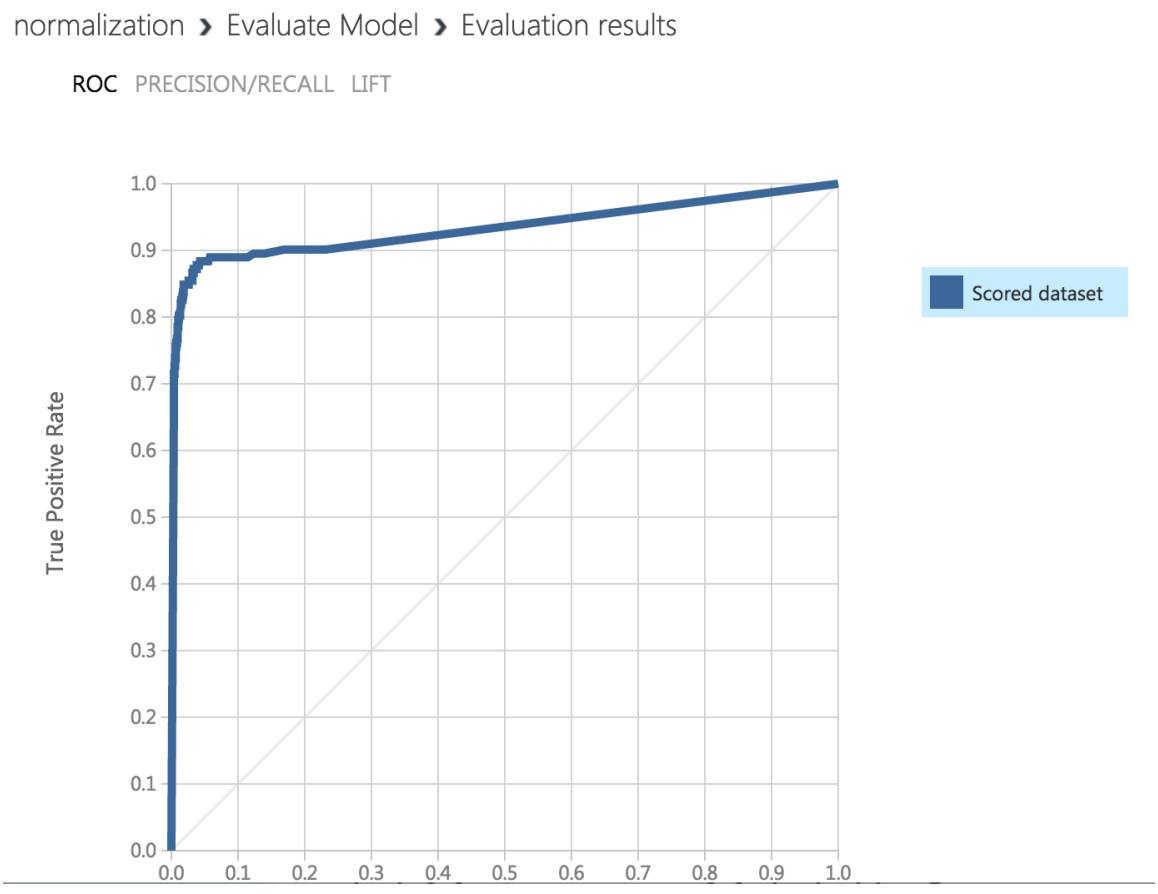
We used the Two-Class Boosted Decision Tree. This tree was selected as it gave the highest Accuracy Rate of 96.3% as the R Square values are close for a few models.

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
142	30	0.963	0.899	0.5	0.927

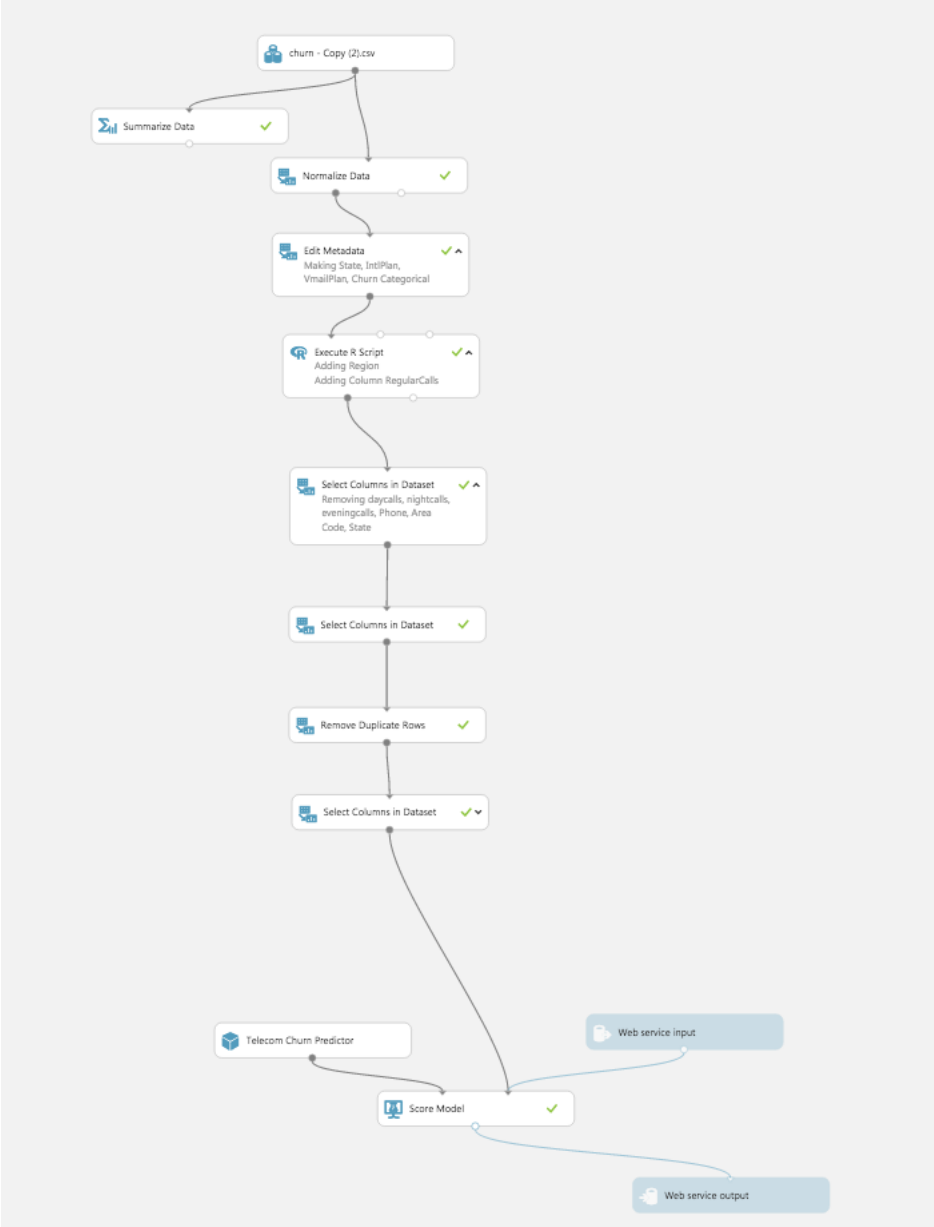
False Positive	True Negative	Recall	F1 Score
16	1062	0.826	0.861

Positive Label	Negative Label
True.	False.

The ROC Curve is as follows:



6. Final Azure Pipeline



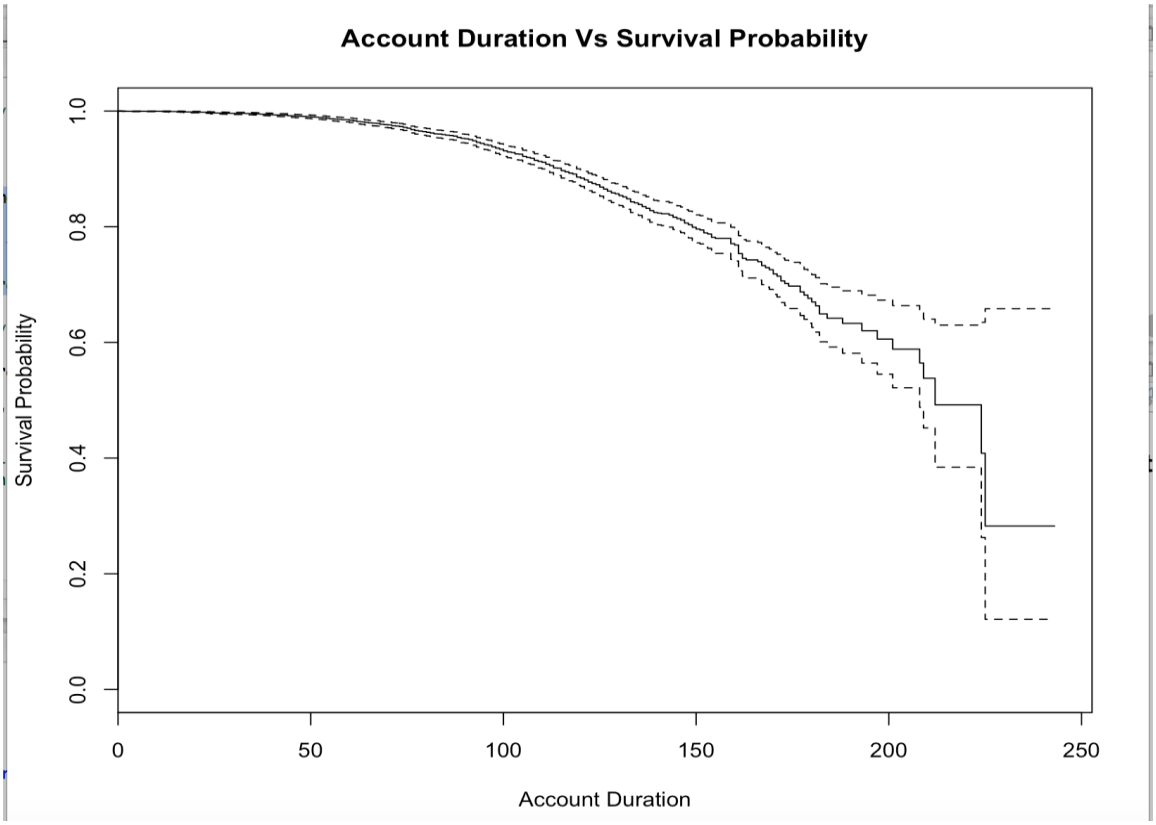
7. Survival Analysis

Conventional statistical methods (e.g. logistics regression, decision tree, and etc.) are very successful in predicting the causes for customer churn. However these methods

cannot predict when customers will churn, or how long the customers will stay with. We use survival analysis for that.

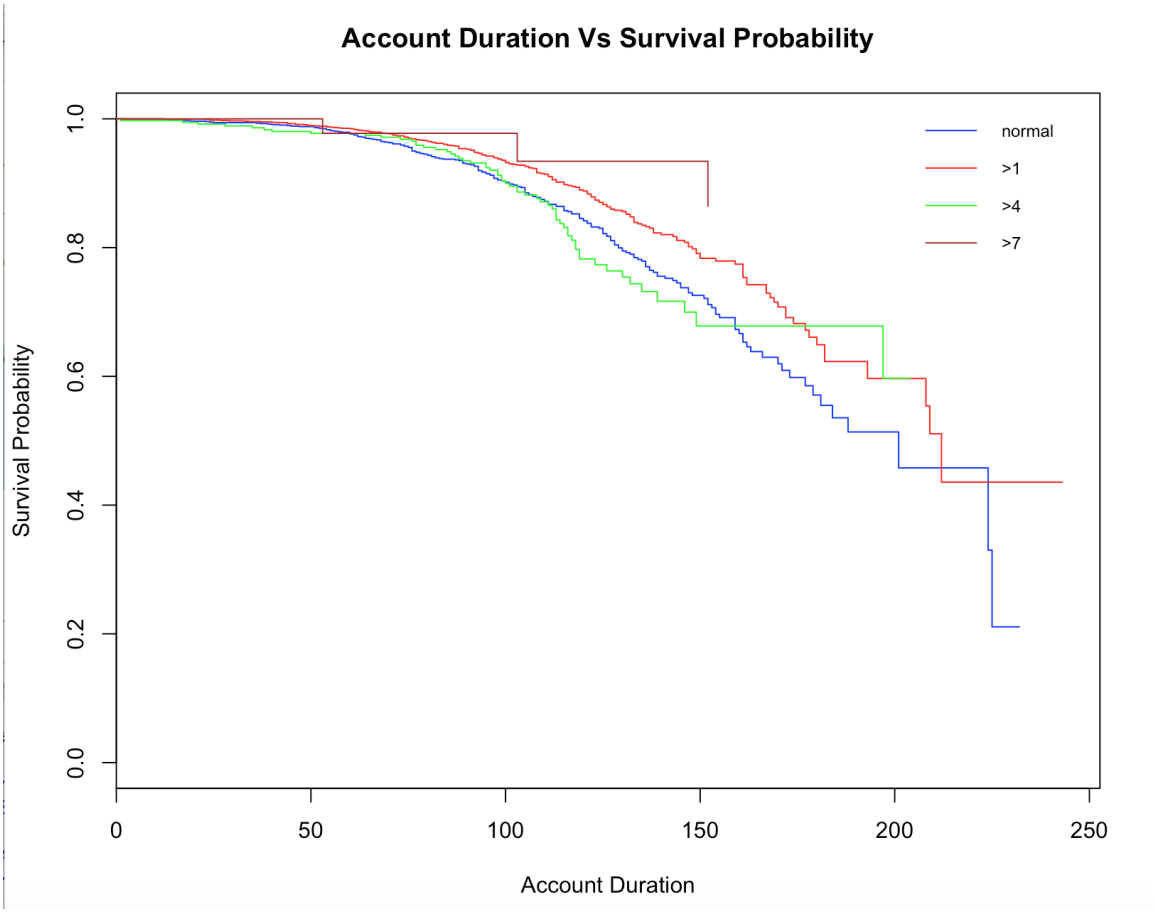
Survival analysis id the branch of statistics that deals with analysis of time duration, in this case we analyze the time duration of customers with the current service provider to predict if the customer will churn.

For our purposes, we use the account duration as our time variable. For the dependent covariates, we use the top 3 predictors of churn from the boosting model.



To make this plot more informative, it would be helpful to take a parameter directly causing churn and vary it to see the effect on survival. Lets consider the number of calls to customer care as the variable and analyze the effect of increases in it on the survival.

Lets evaluate the survival probabilities of survival when we have at least 1, 2, 3, 4 and 5 calls to customer service.



From the graph you can see that with the number of increased customer calls will lead to customer leaving the service provider sooner.

8. Conclusion

In conclusion we have selected the Two-Class boosted Decision Trees as it gives us the best Accuracy even though the R Square value might not be the highest as it is important for the company to correctly identify the customer who are going to leave their services.

I. High correlation between Daytime call charge and Churn to deal with this the company may reduce daytime call tariffs.

   Other service providers maybe providing better rates therefore customer are leaving.

II. There could be a problem in the International Plan as very few people are taking the plan and the people who are taking it are leaving the service provider.

III. Voice Mail Plan is not so popular amongst the people. But the people who have taken it seem to be happy with it therefore, the company could advertise more on the voicemail plan.

IV. Customers Account Duration doesn't seem to have a great impact on the churn Rate, the company might not have any benefits for loyalty customers.

V. The customer service calls should be less, as that is an important factor for churn.

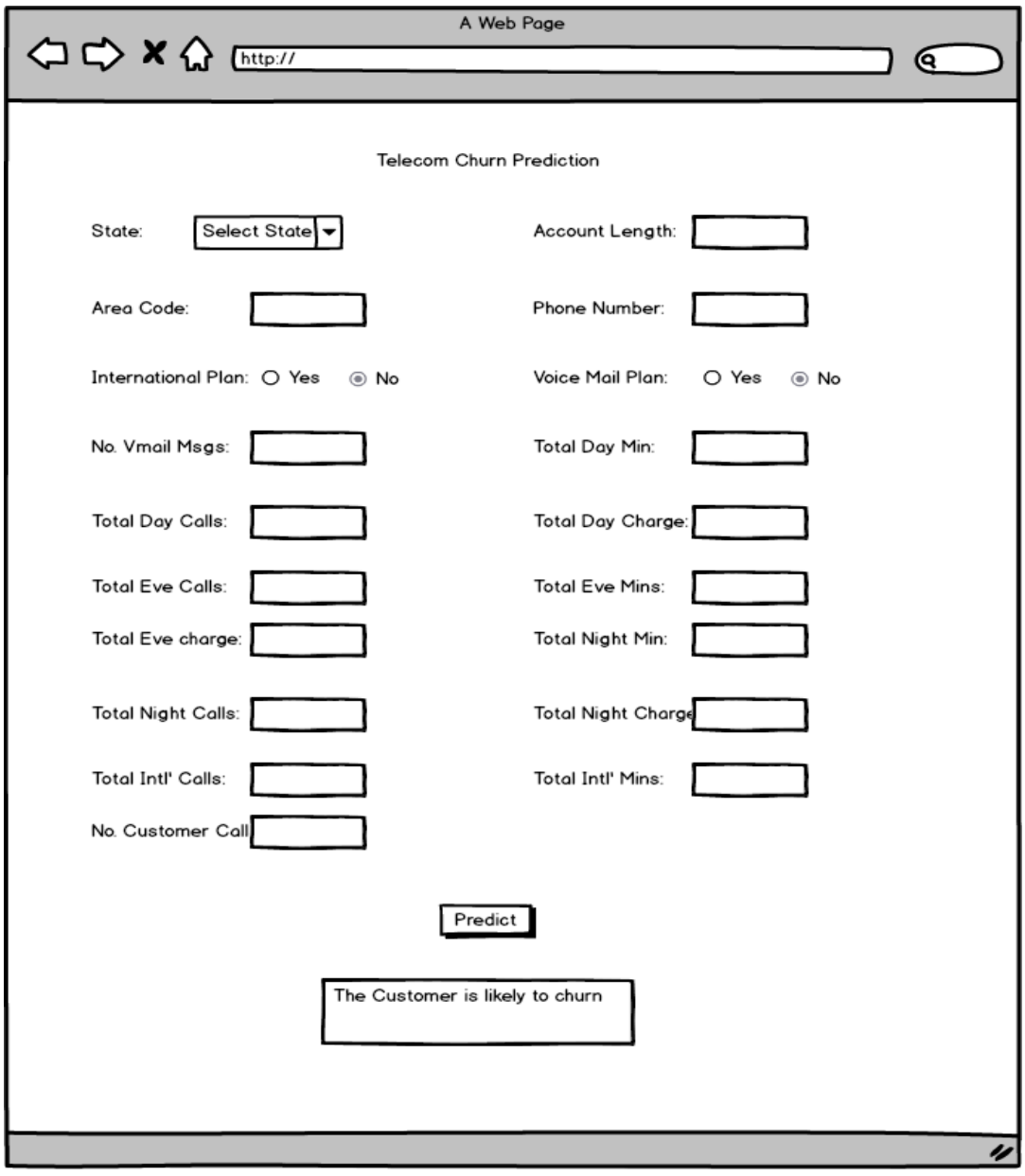
VI. On Survival Analysis we found that the chance of a customer leaving the company increases as 100 to 120 months of owning the account.

Web Application development

In designing the web we finalized the Visual studio for development as Azure itself gives good support to visual studio and compatibility is high .

Design:

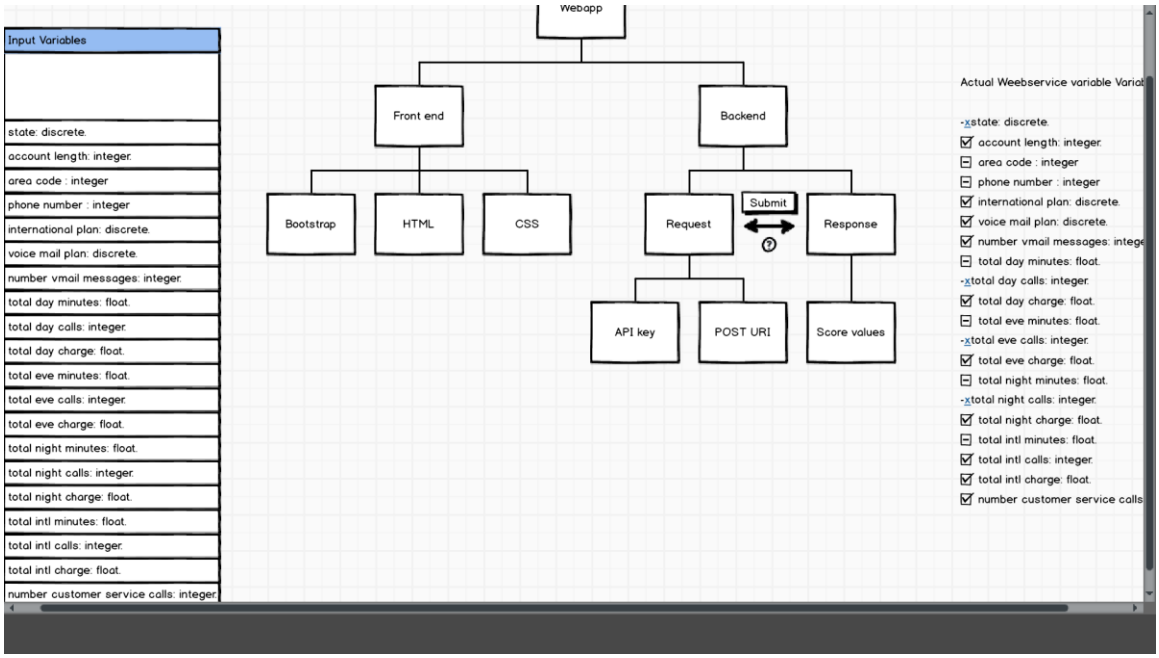
Front End:



Backend and Integration:

It has flow diagram and two sections. The left section has all the variables that will be given by customer and right section includes what different variable we will use and how. the checked Box are variables taken as it is and [-] boxes are not taken into

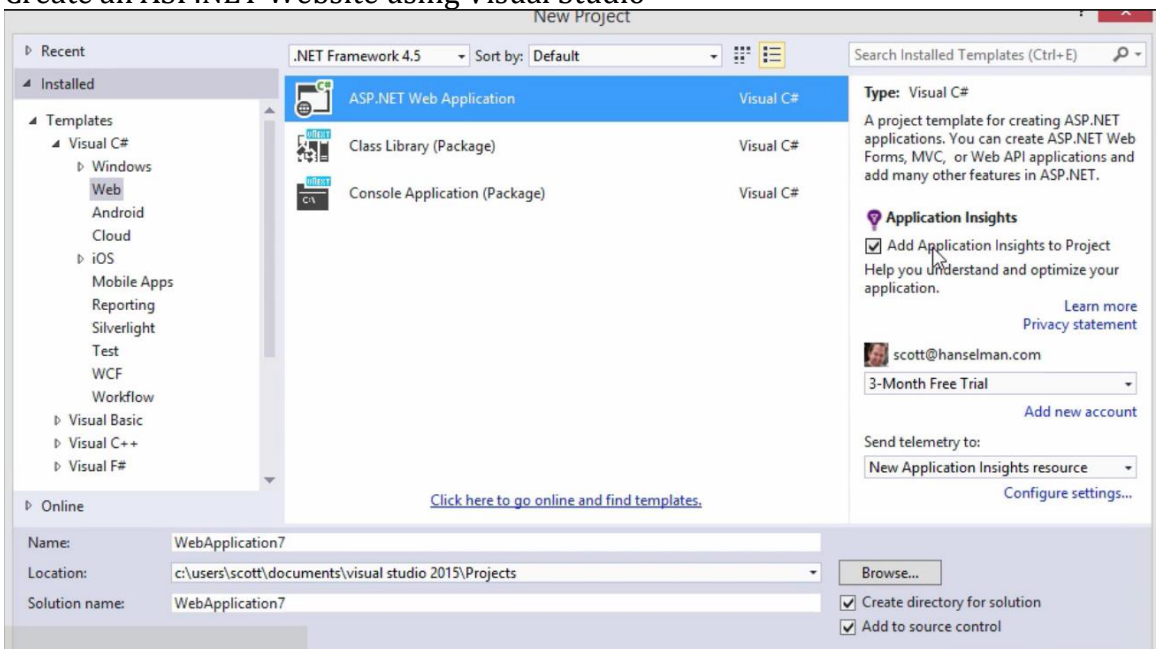
consideration as they are highly correlated and doesn't make impact. [-x] boxes are the one which we have considered for feature selection create new features.



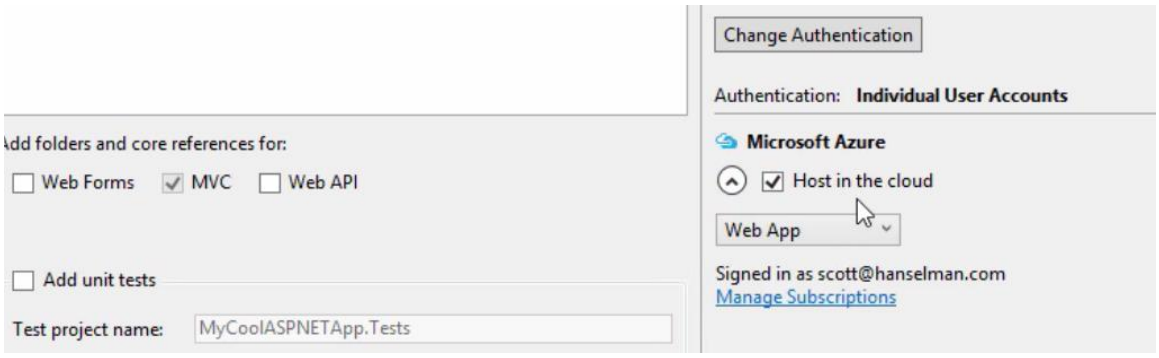
Development and Deployment:

After deploying the web service on azure we need API Key, POST URI and C# code that azure generates for reference and fetching the data using the web service.
We followed below steps to create the web app: -

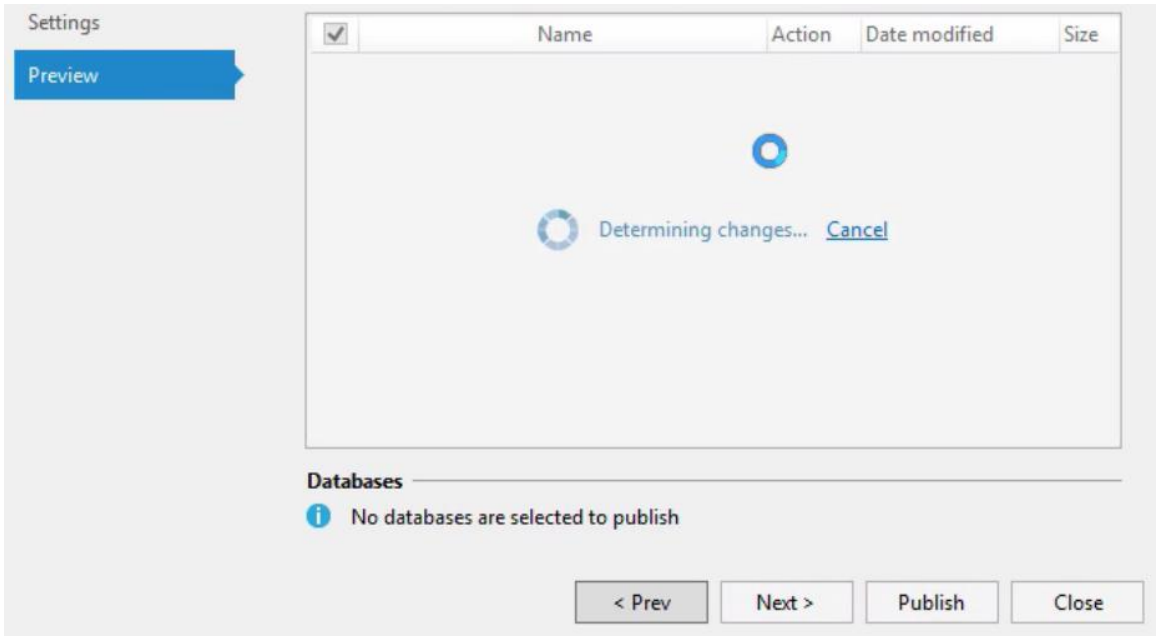
- Create an ASP.NET Website using Visual Studio



- Select option for Host in cloud:

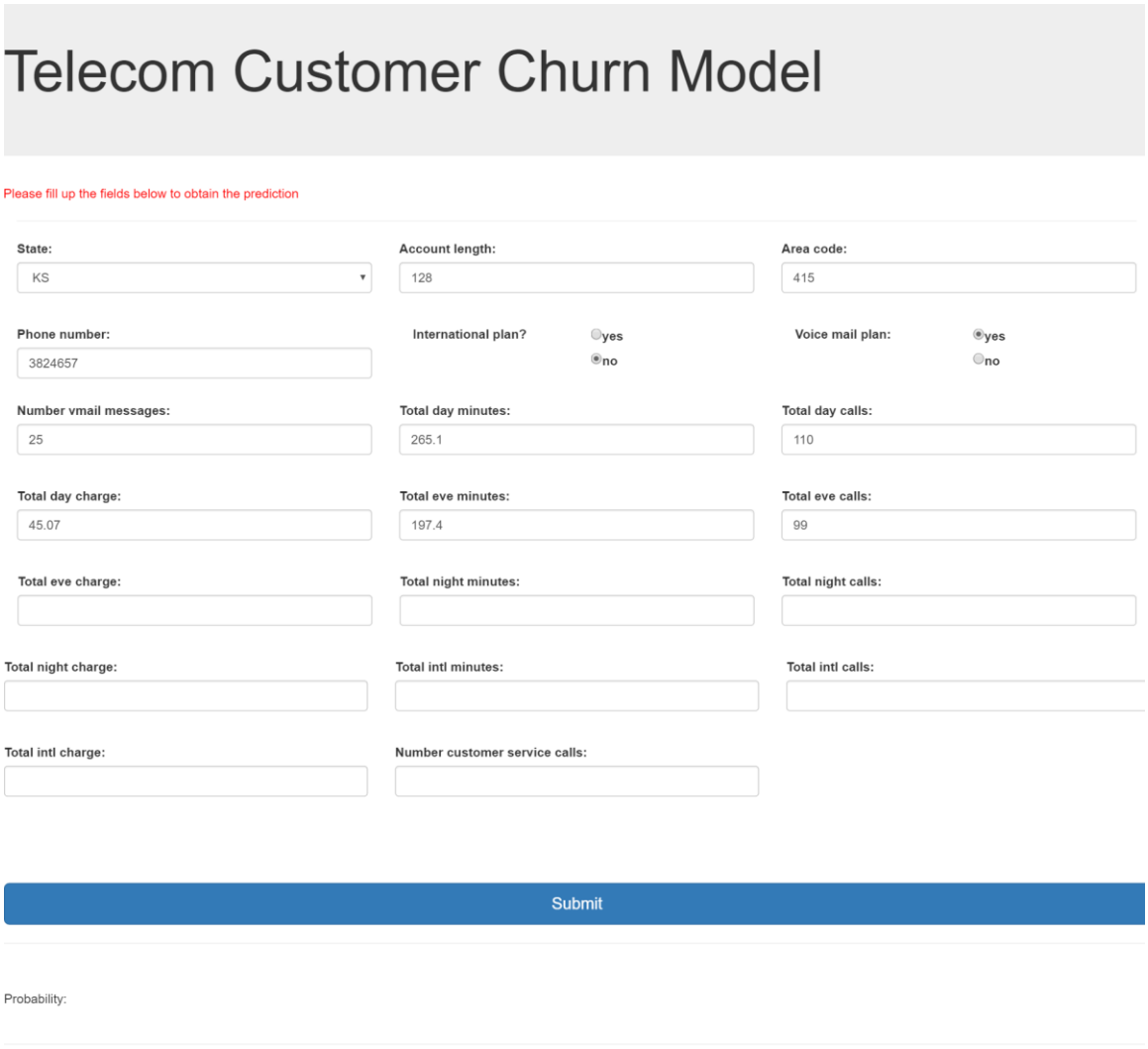


- Develop the web-project and publish it



- Look at the given website path to access the web application: http://adsteam1.azurewebsites.net/webform1.aspx

Snapshots of the Webapp:



Use cases:
There are two output possible for the customer churn, TRUE (customer will churn) and FALSE (customer will not churn). Below are some sample input and expected output :

State	OR	MD
AccountLength	59	135
AreaCode	408	408
Phone	353-3061	383-6029
IntlPlan	no	yes
VMailPlan	yes	yes
VMailMessage	28	41
DayMins	120.9	173.1
DayCalls	97	85
DayCharge	20.55	29.43
EveMins	213	203.9
EveCalls	92	107
EveCharge	18.11	17.33
NightMins	163.1	122.2
NightCalls	116	78
NightCharge	7.34	5.5
IntlMins	8.5	14.6
IntlCalls	5	15
IntlCharge	2.3	3.94
CustServCalls	2	0
PROBABILTY	**False.**	**True.**

We have also handled the client-side and server side validations for filtering the dataset. Values that doesn't contribute to model are taken as option at the webpage.