

PROJECT REPORT

Predictive Analytics

**By:** Niharika Dobanaboina

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Contents** | **Page no.** |
| 1. | Project Goal | 2 |
| 2. | Overview of Data, including data exploration Analysis | 2-3 |
| 3. | Details of Modelling Strategy | 3-4 |
| 4. | Estimation of Model’s Performance | 5-6 |
| 5. | Insights and Conclusions | 7 |

1. **Project Goal:**

The objective of this project is to help ABC Wireless Inc. identify the customers who are most likely to churn. The primary goal is to develop a model using historical data to predict the customers who might leave the company. These predictions will help ABC Wireless Inc. to implement new business strategies to stop their customers from leaving. Since acquiring a new customer is more expensive than keeping their existing customers, the final interpretation of this project will be beneficial to the company both in terms of revenue and customer satisfaction.

1. **Overview of Data:**

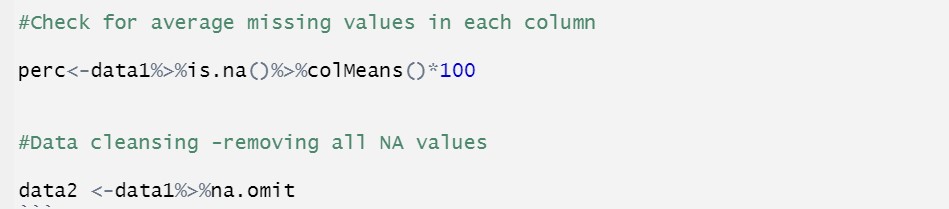
This section of the report contains various analysis techniques applied to the data before developing a model. It includes both *Data Preparation* and *Data Exploration Analysis* used on the dataset.

* ***Data Preparation:***

Initial Training dataset **Churn\_Training.csv** has 20 predictor variables.

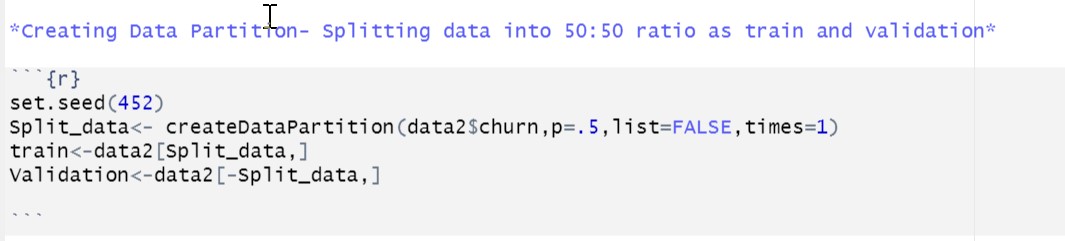
Checked the percentage of null values in each column to identify the columns having a maximum percentage of null values.

Omitted rows with null values:



Partitioned the dataset into Training and Validation sets with the Training set having 50% of the data and the Validation set with 50% of the data.

The reason behind splitting the data was to better understand the performance of the model as to how good will the model be in predicting the probabilities on the Test dataset:



* ***Data Exploration:***

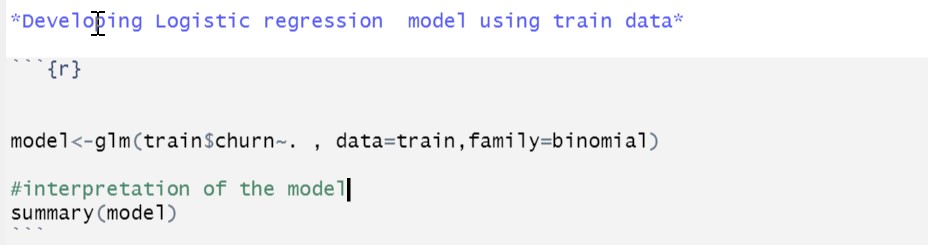
Converted the churn variable into factors datatype with 2 levels of “yes” and “no” values.



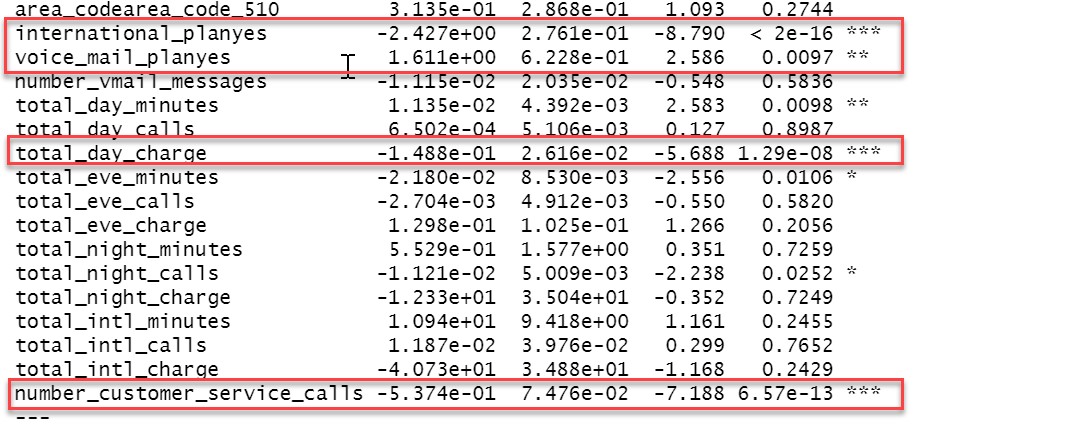
**3. Modeling Strategy:**

As our dataset is a mix of both categorical and numerical data, we considered **Logistic Regression** to be the best one to develop a model to predict the **churn variable** which also happens to be a categorical variable.

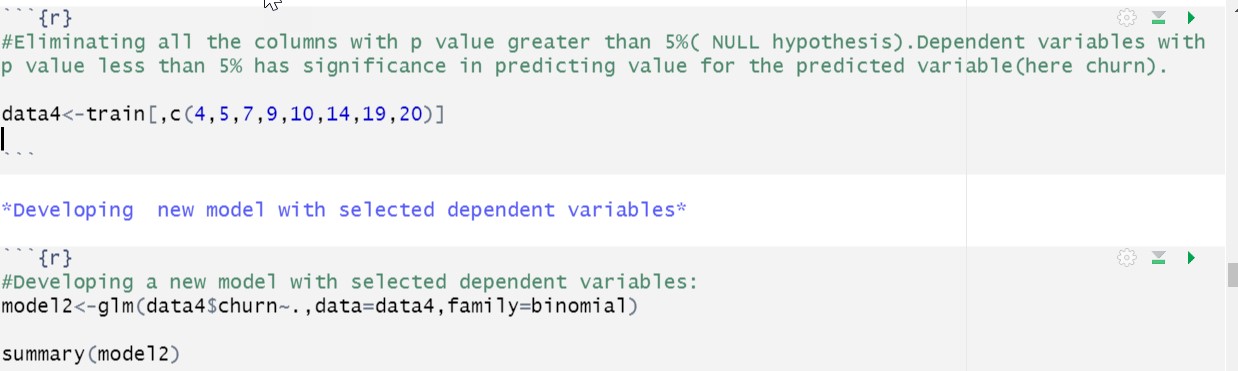
Though we had the option to convert the “yes” and “no” values of the churn variable to 1s and 0s to use simple linear regression, we did not proceed with that idea as there were chances of predicted probabilities to being less than 0 or greater than 1.

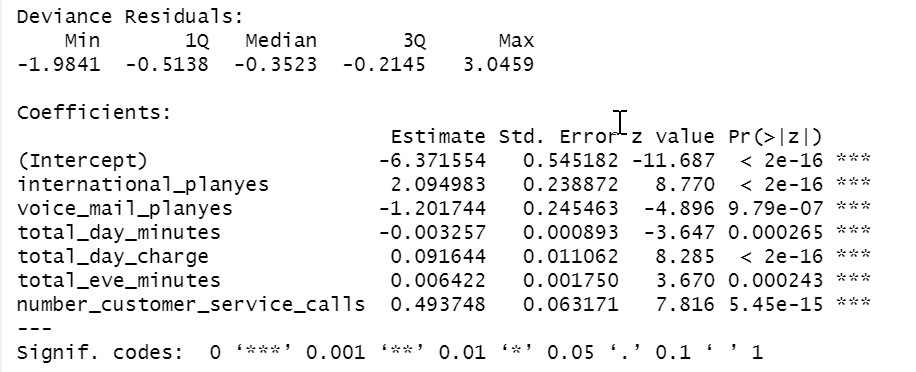


After running the model successfully, we identified the variables which have a significant relationship with the predictor variable by looking at the “p” value. We considered the threshold to be 0.5 and all the variables with a p-value less than 0.5, were subsetted into a new data frame to proceed with further analysis.



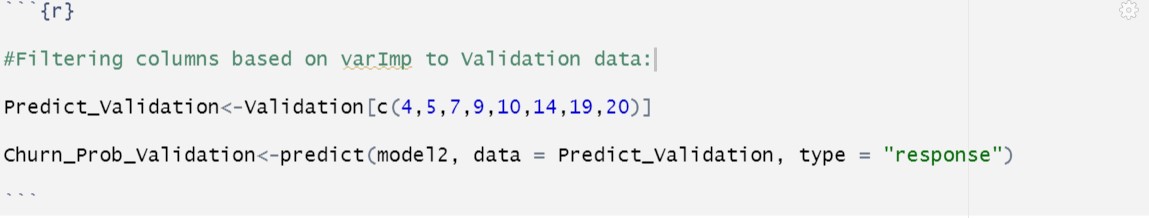
Developed a new model with selected variables and cross-verified the p-value. As all these variables hold a significant relationship with the independent variable, we considered this to be the best model to predict the churn value.





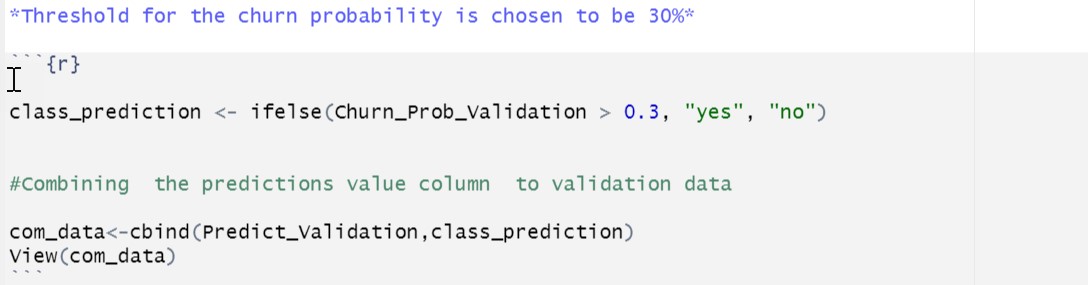
**4. Estimation of Model’s Performance:**

To know how well the model works on unseen data, we ran this model on the validation dataset.

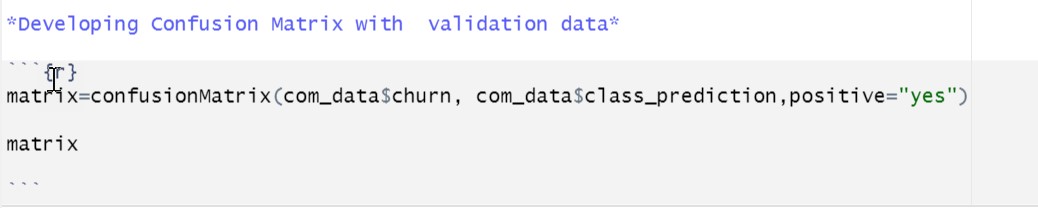


After predicting the probabilities on the validation set, we converted those predicted values into “yes” or “no” considering the threshold to be 0.3. All the predicted values having percentages greater than 0.3 were assigned “yes” and the rest were assigned a “no” value.

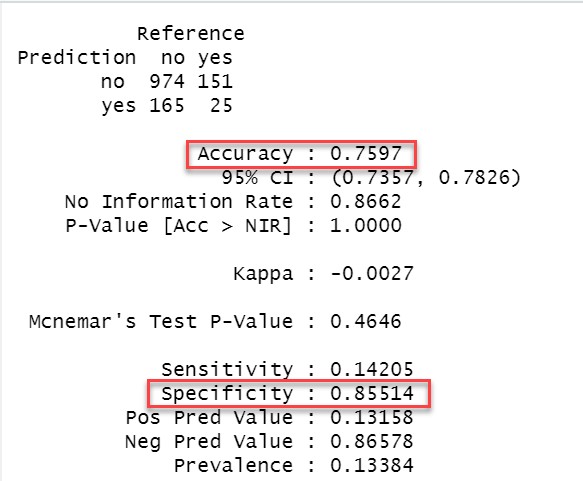
This cutoff of 0.3 was decided based on the business revenue aspect. It would be a loss of cost for a company even if a single customer churns. So, it is safer to identify the greatest number of customers who might leave, in this case, even if that means that the probability of churning is as low as 30%.



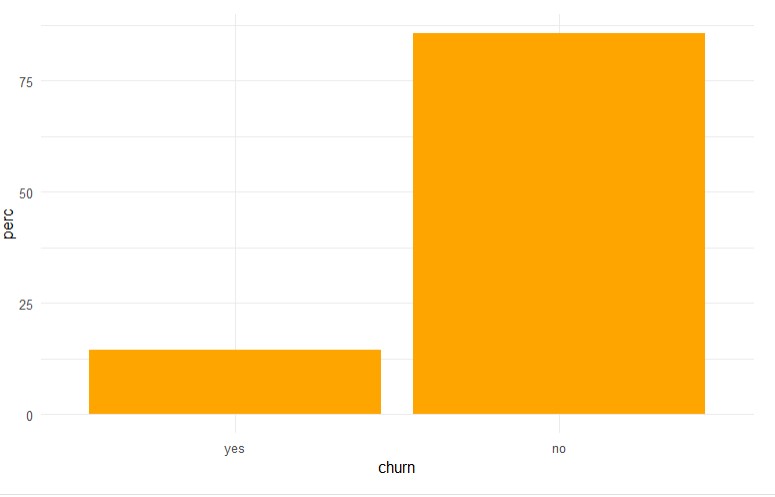
Appended these values to the training dataset and ran a confusion matrix:



Interpretation of the Model’s Performance based on the results of confusion matrix:



The accuracy of the model is 75.97%. As the below plot indicates, the data is unbalanced, and this factor has an impact on the Accuracy.

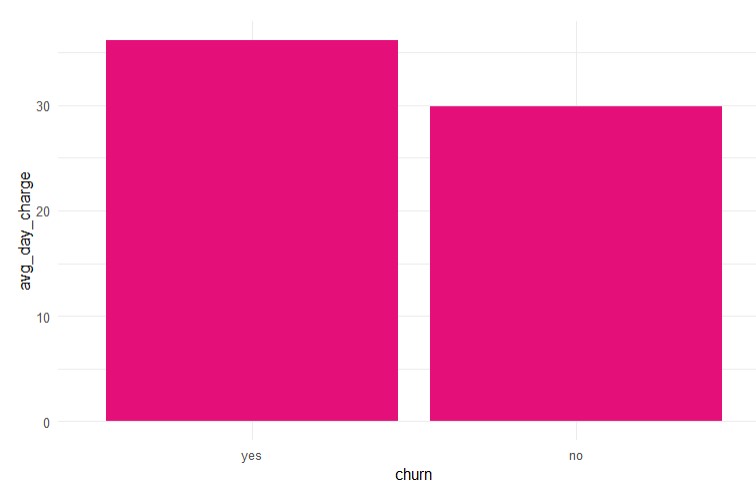
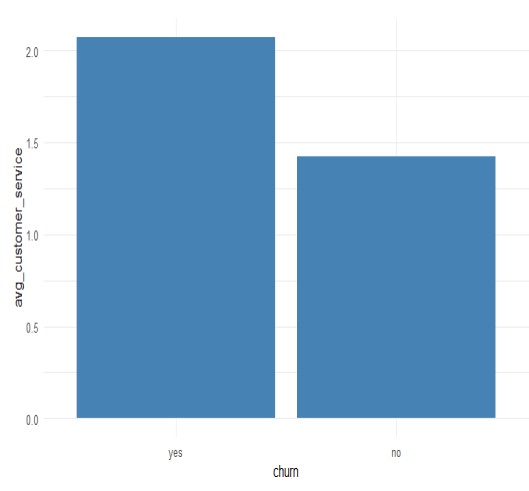


Another performance metric we considered is Specificity. The True Negative Rate of this model is 85.55%. The reason for considering this performance metric is that if we identify any customer as the one who is going to churn out wrongly, then the company will invest in strategies for that customer who was anyway going to stay. This will be an additional loss to the company as they are spending unnecessarily on such customers. Hence, it becomes important to us to not identify any customer to be churning out falsely.

**5. Insights and Conclusions**

Through our analysis, we tried to identify a few reasons as to why a customer might want to leave the company.

Based on the below plot, the customers who made the greatest number of customer service calls are to be the ones who are most likely to churn. These could be the customers who might have called the most times to raise their concerns. So, the company should target customers who are calling them frequently and solve their problems as a priority. This will also improve customer satisfaction.



The above plot indicates that the customers who are spending more on an average every day are the ones who might churn out. These customers could probably think that they are spending more on this network and if they are getting better deals at another company, they might consider churn out. Therefore, the company should target the customers who have taken the largest spending plan per day and offer discounts to them.

**Conclusion:**

Customers who were most likely to churn were predicted on the Test data. These predictions will help ABS Wireless Inc. to retain their existing customers while also improving their revenue, and customer relationships.