Using Linear Discriminant Analysis to Predict Customer Churn

datascience.com/blog/predicting-customer-churn-with-a-discriminant-analysis

In a competitive world, the key to business success is to understand enough about your customers' behavior and preferences so that you can provide a personalized service to both your prospective and existing customer base. Using customer behavior analytics techniques, you can predict how a customer will respond to a business situation, thereby giving you the information you need to approach them. These techniques can be broadly grouped into two buckets:

1. **Regression problems**: The response to this type of problem is a numerical value; for example, based on customer demographics and past purchasing patterns, how much will the customer spend in a particular instance?
2. **Classification problems**: The response to this type of problem is categorical; for example, based on the customer demographics and his or her past experience with the product/service, will the customer stay or leave the product/service?

Predicting whether a customer will stop using your product or service is an important component of customer behavior analytics called churn prediction. In this post, I will analyze two aspects of churn prediction:

1. For the given scenario, which factors are mainly responsible for customer churn?
2. Given the data pertaining to a new customer, how do you predict whether the customer will churn or stay, and what are the associated probabilities?

The Business Problem: Predicting Churn at a Telecom Service Provider

The telecom business is challenged by frequent customer churn due to several factors related to service and customer demographics. The dataset we'll use in our analysis includes a list of service-related factors about existing customers and information about whether they have stayed or left the service provider. Our objective is to understand which of the factors contribute most to customer churn and to predict which customers will potentially churn based on service-related factors.

About the Dataset

The dataset used for the analysis can be downloaded here. It consists of information for 5,000 customers and includes independent variables such as account length, number of voicemail messages, total daytime charge, total evening charge, total night charge, total international charge, and number of customer service calls.

The dependent variable in the dataset is whether the customer churned or not, which is indicated by a 1 for "yes" and 0 for "no."

What is Discriminant Analysis?

In order to uncover which variables are responsible for churn and predict whether a customer will churn or not, we will use discriminant analysis.

Discriminant analysis is a segmentation tool. It segments groups in a way as to achieve maximum separation between them. This technique makes use of the information provided by the X variables to achieve the clearest possible separation between two groups (in our case, the two groups are customers who stay and customers who churn). Below is our formula:

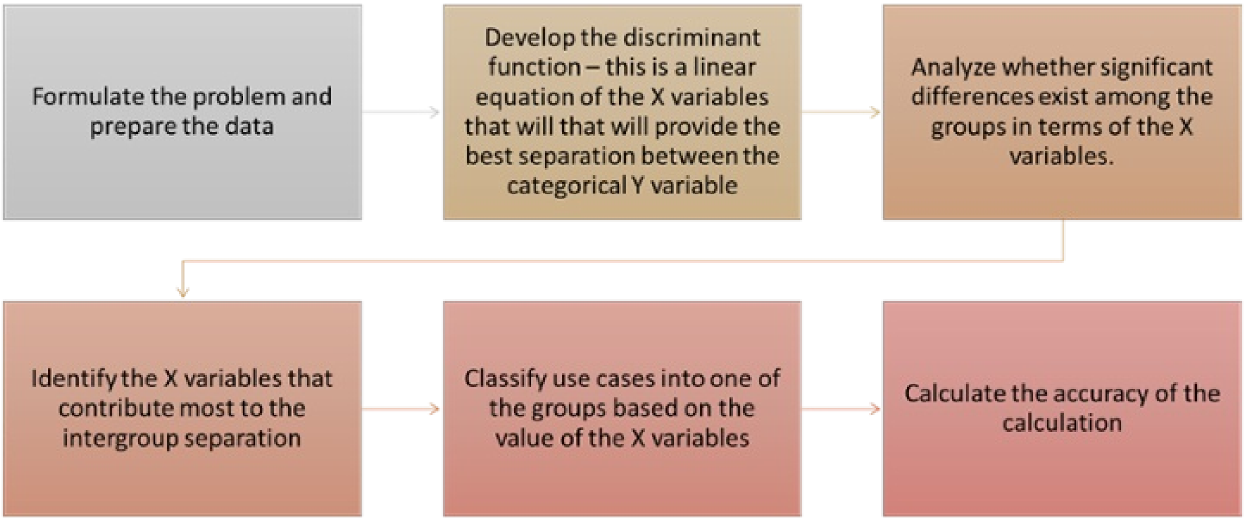
D= b0 + b1X1 + b2X2 + .. bnXn

Here, D is the discriminant score, b is the discriminant coefficient, and X1 and X2 are independent variables.

The discriminant coefficient is estimated by maximizing the ratio of the variation between the classes of customers and the variation within the classes. In other words, points belonging to the same class should be close together, while also being far away from the other clusters.

Discriminant Analysis: A High-Level Approach

Below, I've broken down the steps we will be following in order to predict customer churn, starting with data preparation and ending with validating the accuracy of our model.


Formulate the problem and prepare the data
Develop the discriminant function – this is a linear equation of the X variables that will that will provide the best separation between the categorical Y variable
Analyze whether significant differences exist among the groups in terms of the X variables.
Identify the X variables that contribute most to the intergroup separation
Classify use cases into one of the groups based on the value of the X variables
Calculate the accuracy of the calculation

Let's get started.

Formulating the Problem and Preparing the Data

To begin, we need to read the CSV file in R and convert the target variables into categorical variables. For the purpose of discriminant analysis, the Y variable has to be a categorical variable (meaning it is one of a limited number of possible values).

```
setwd("C:/Users/Sowmya CR/Google Drive/datascience_blog")

data=read.csv("churn2.csv")
```

```
data$churn=factor(data$churn)

str(data)
```

Now that we've converted our target variables, let's find the baseline churn rate for the dataset:

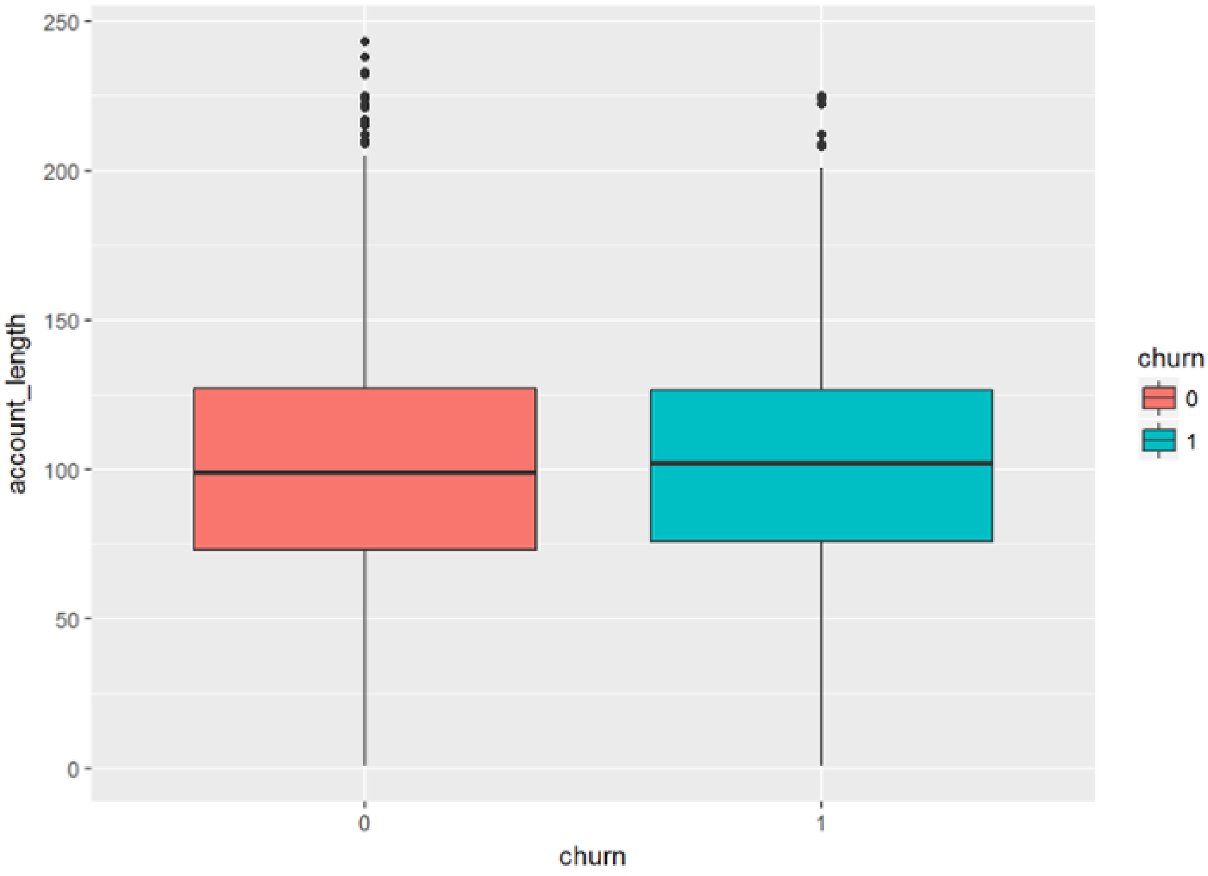
```
prop.table((table(data$churn)))
```

As you can see here, the baseline churn rate is 14.14%, which is an indicator that the dataset is unbalanced — there are more 0s than 1s. Subsequently, this will have an impact on how we interpret model performance measures.

Exploratory Data Analysis

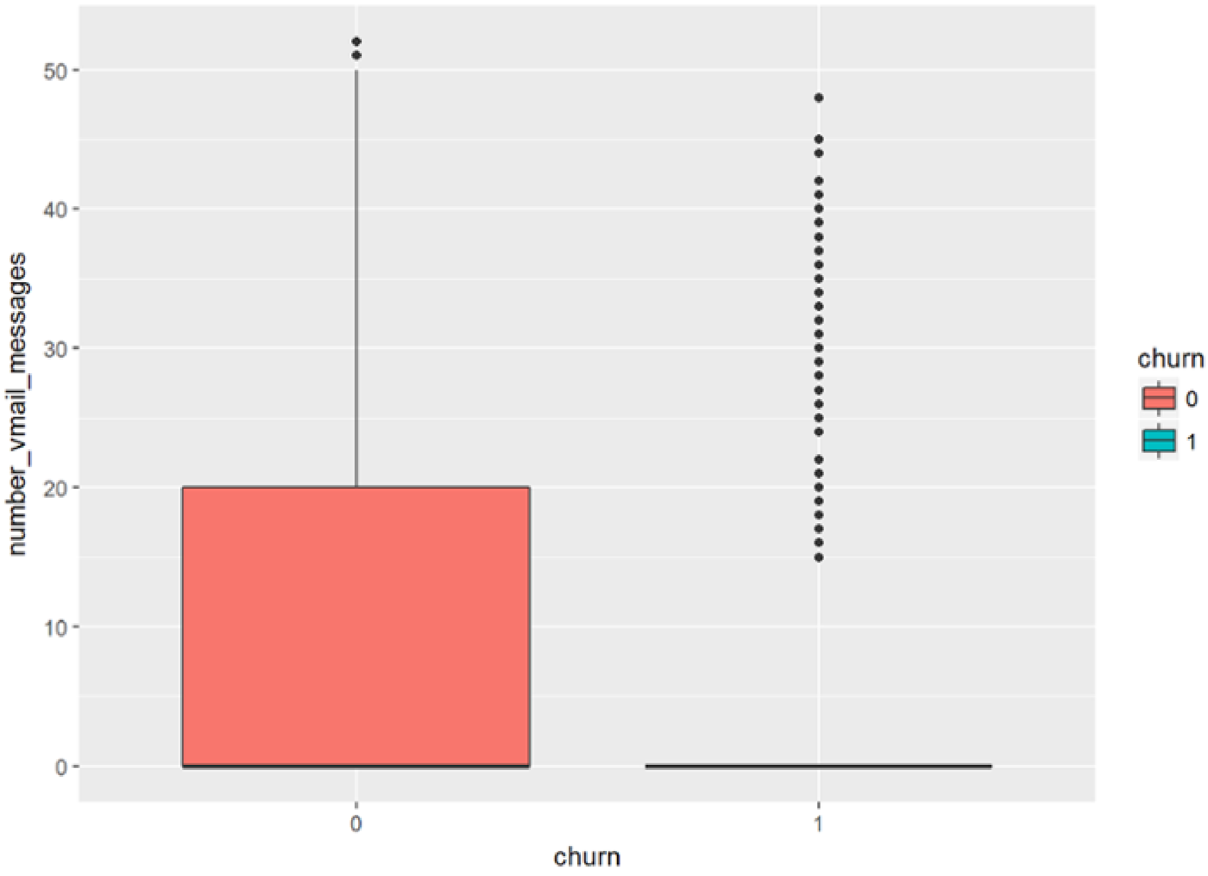
Using ggplot2, I will explore our dataset and determine the impact our variables have on churn, starting with account length.

```
ggplot(data, aes(x=churn, y=account_length, fill=churn)) +
geom_boxplot()
```



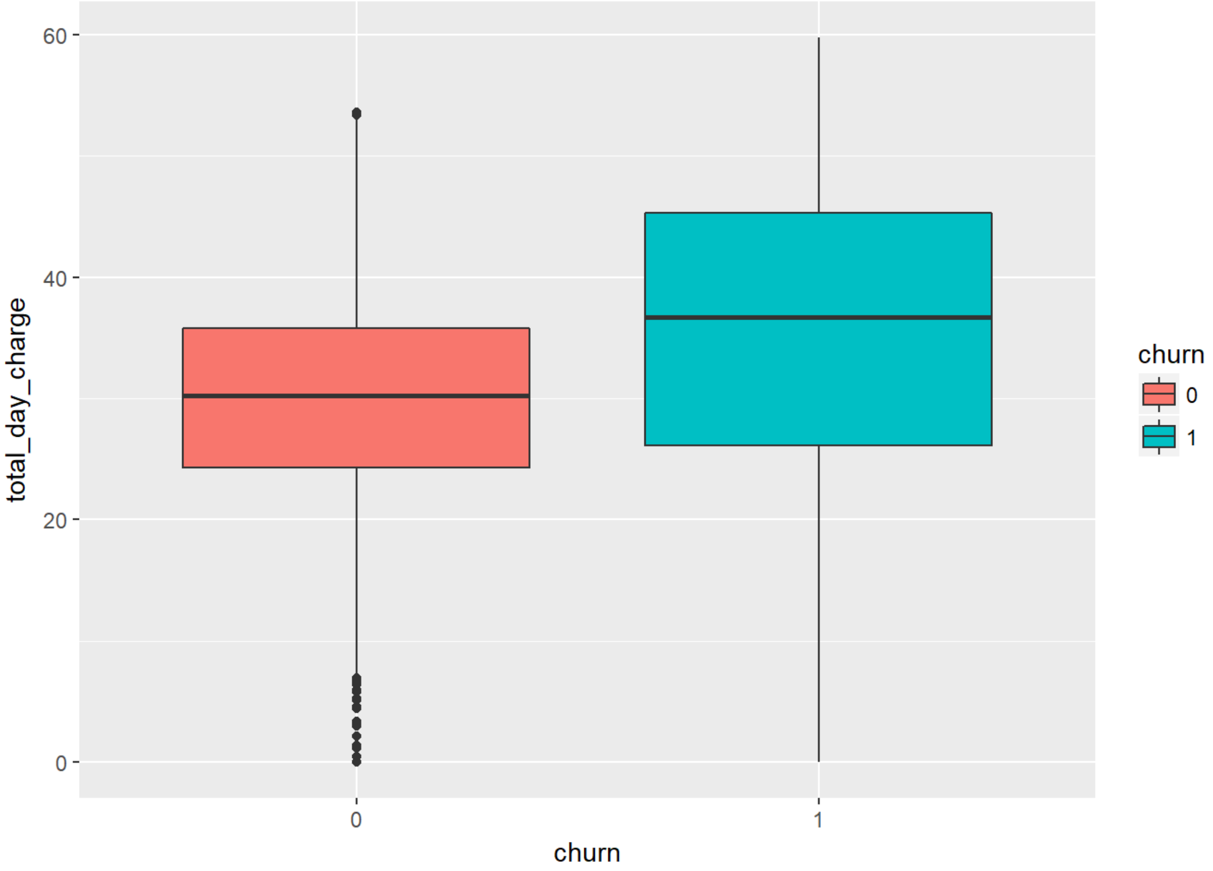
As you can see in our box plot, account length does not seem to have an influence on customer churn. Now, let's look at voicemails:

```
ggplot(data, aes(x=churn, y=number_vmail_messages, fill=churn)) +
geom_boxplot()
```



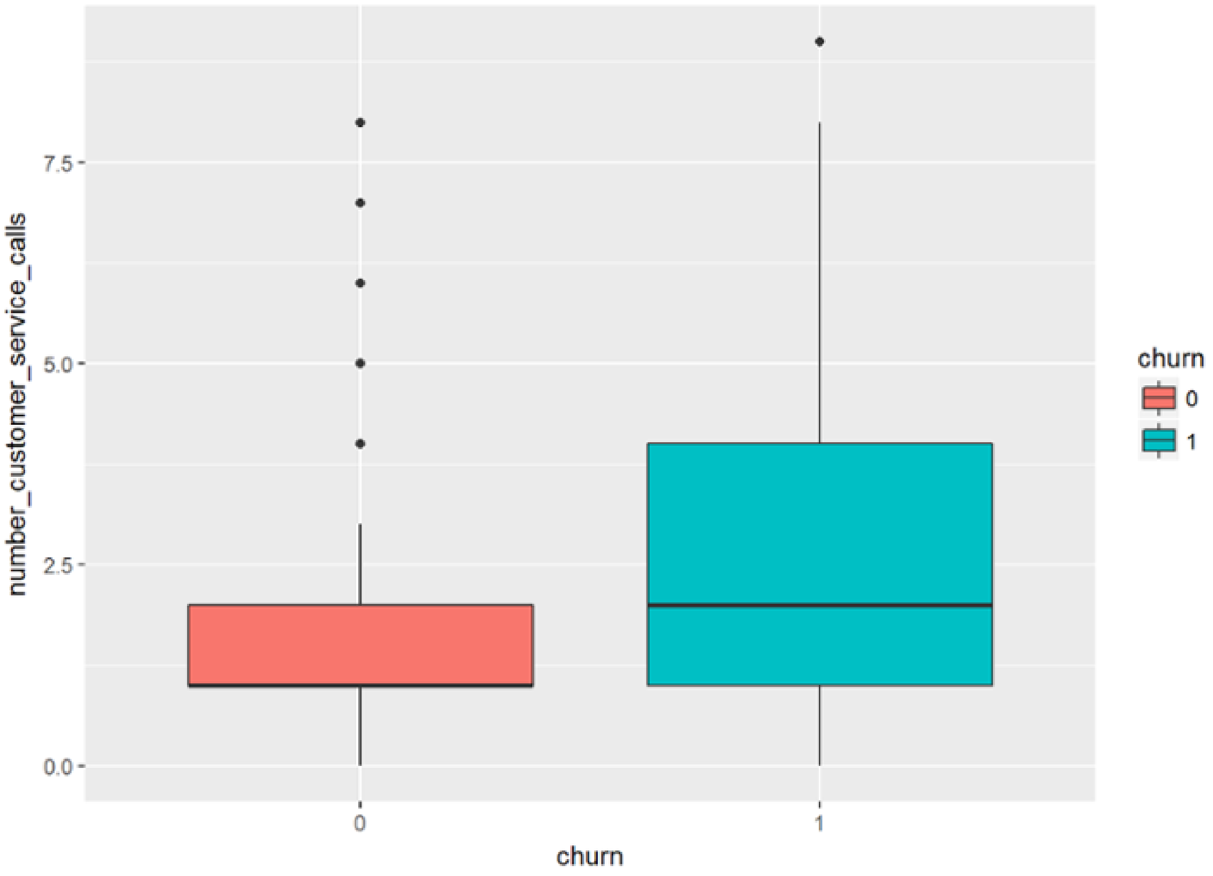
The number of voicemail messages seems to be lower for customers who churn compared to those who don't.

```
ggplot(data, aes(x=churn, y=total_day_charge, fill=churn)) +
geom_boxplot()
```



Generally, tariffs for day, evening, night, and international calls are higher for customers who churn compared to those who don't. This could indicate that customers who churn are not happy with the amount of money they are paying for their plan.

```
ggplot(data, aes(x=churn, y=number_customer_service_calls, fill=churn)) +
geom_boxplot()
```



Here, we see that the number of customer service calls made to customers who churn is relatively high. This indicates that customers who have churned have tried contacting customer service — but might have not received a satisfactory resolution to their issue.

Using the assumptions we've derived from our exploratory data analysis, we will dive deeper into our variables using discriminant analysis.

Since the data have variables that are in various dimensions, it is advisable to scale them so the range of each variable does not have an influence on the discriminant coefficients. After the dataset is scaled, we will split the dataset into training and test datasets. We will build the model based on the training dataset and test the model performance using the test dataset.:

```
x <- subset(data,select=c(1:7))
scaled_x=scale(x)
data1=cbind(data[8],scaled_x)



library(caTools)

set.seed(123)
split = sample.split(data1$churn, SplitRatio = 0.7)
traindata = subset(data1, split == TRUE)
testdata = subset(data1, split == FALSE)


prop.table((table(traindata$churn)))
prop.table((table(testdata$churn)))
```

Step One: Test the Significance of the Discriminant Function Using MANOVA

To test the significance of the discriminant function, we will use multivariate analysis of variance (MANOVA). Here's the approach we will take:

1. The null hypothesis of MANOVA is that all the means of the independent variables are equal, which implies that the independent variables are not differentiators of the group.
2. The alternative hypothesis is that at least one independent variable has a different mean or, in other words, a significant differentiator.

```
head(traindata)
```

```
X1=cbind(as.matrix(traindata[,2:8]))
Y1=as.vector(traindata[,1])
Manova=manova(X1~Y1)
summary(Manova, test = "Wilks")
```

As you see above, the Wilks' lambda for MANOVA is closer to 1, indicating that the extent of discrimination in the model is relatively low. But the *p*-value is highly significant, indicating that the null hypothesis cannot be accepted. This implies that the discriminant model is highly significant.

Step 2: Develop the Fisher Discriminant Function

Now we need to identify a combination of features that separates our customers that are likely to churn from those who are not. I'll be using the packages DiscriMiner and MASS to do so.

```
library(DiscriMiner)
library(MASS)
```

```
discPower(X1,Y1)
```

```
desDA(X1,Y1)
```

Since there are only two values for the Y variable, there is only one discriminant function DF1. The coefficients of the discriminant function can be seen below:

If we sort the X variables by descending order of the coefficients, we are able to understand the influence of each X variable on differentiating the Y variable:

```
constant 0.02478
number_customer_service_calls 0.70106
total_day_charge 0.62922
total_eve_charge 0.28724
total_intl_charge 0.25025
total_night_charge 0.16
account_length 0.05173
number_vmail_messages -0.30702
```

In terms of magnitude, the number of customer service calls has the most impact, whereas the account length has the least impact. Number of voice mail messages has a negative sign indicating that it has a negative impact on churn. In other words, as the number of voice mail messages increases, the probability of churn decreases.

Step 3: Differentiation of Individual Independent Variables and Wilks' Lamba

The *p*-value is not significant for account length, but it highly significant for the other X variables. This indicates that all X variables except account length are excellent predictors in terms of differentiating our customer groups.

Step 4: Correlation Between Discriminant Function and Independent Variable

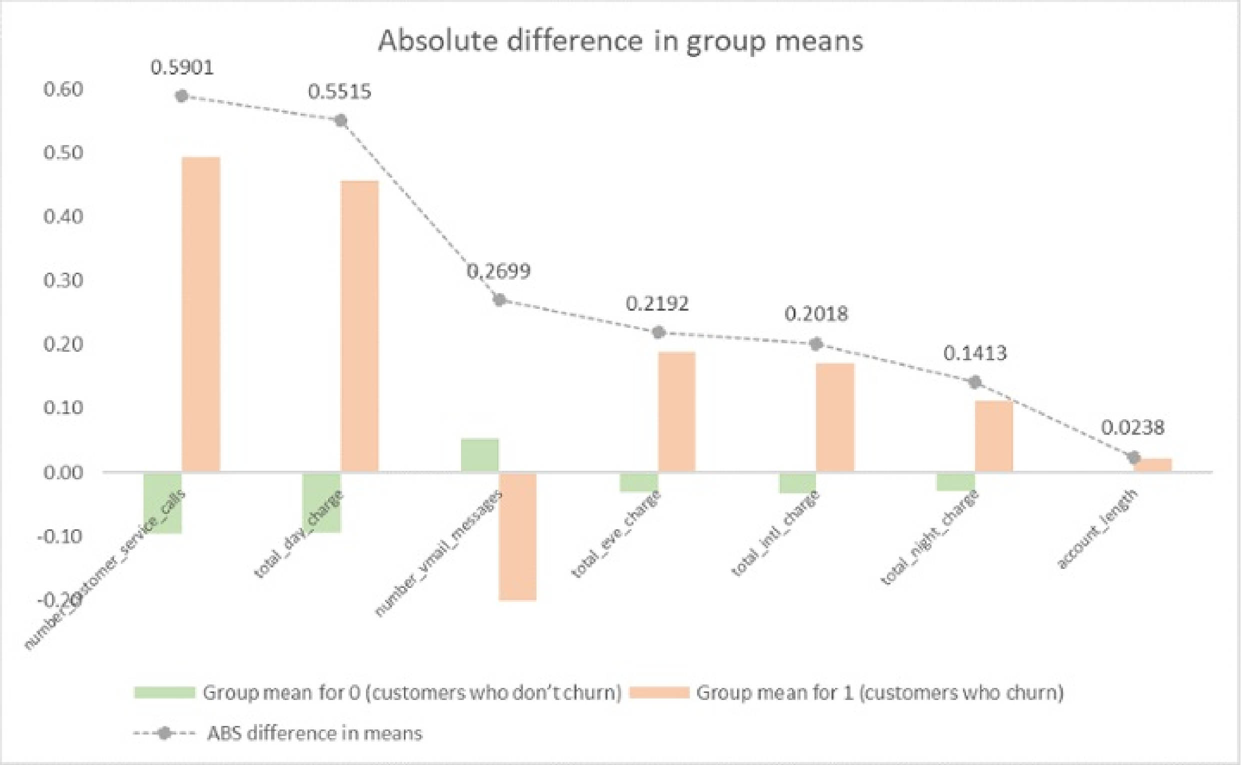
The relative importance of the X variables can be determined by the correlation ratio of each of the variables. The number of customer service calls emerges as the most significant variable in its ability to differentiate the groups.

Step 5: Classify Records Based on Discriminant Analysis of X Variables and Predict Y Variables for the Test Set

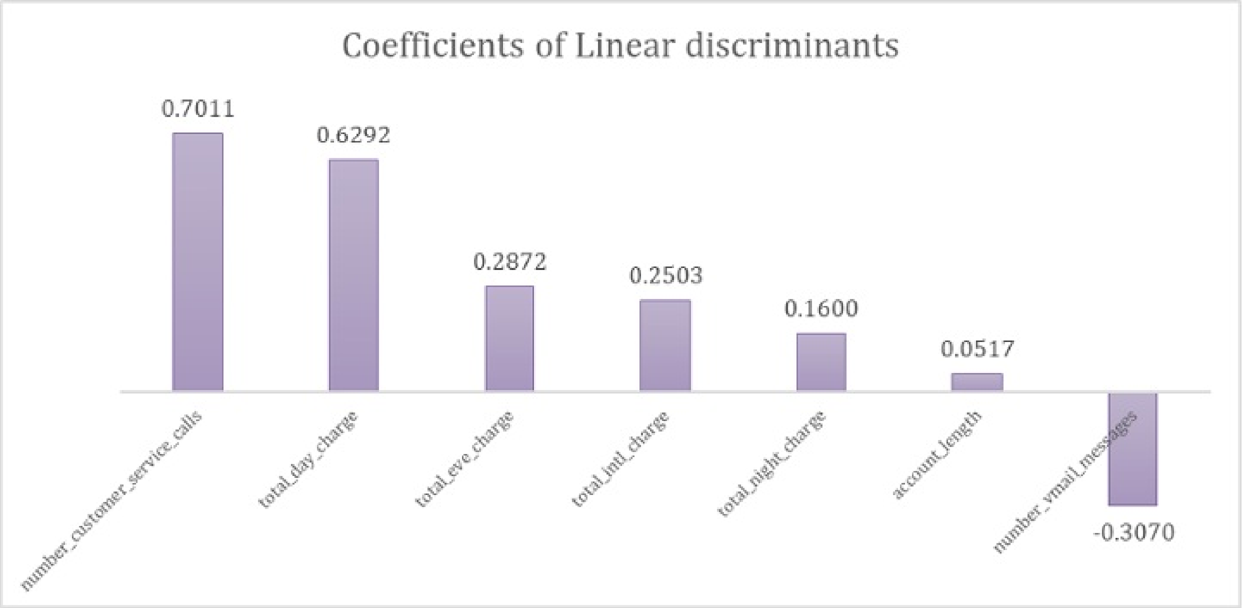
Since the discriminant model is significant, we will use it to classify records as belonging to either customers who will churn or those who will not churn depending on the X variables. We will use the lda() function in R to classify records based on value of X variables and predict the class and probability for the test set.

```
sublda=lda(churn~.,data = traindata)
sublda
```

Let us review the output from the lda() function:



The difference in group means is highest for number of customer service calls and lowest for account length. This gives us insight into which factors contribute most to the discrimination between the groups.

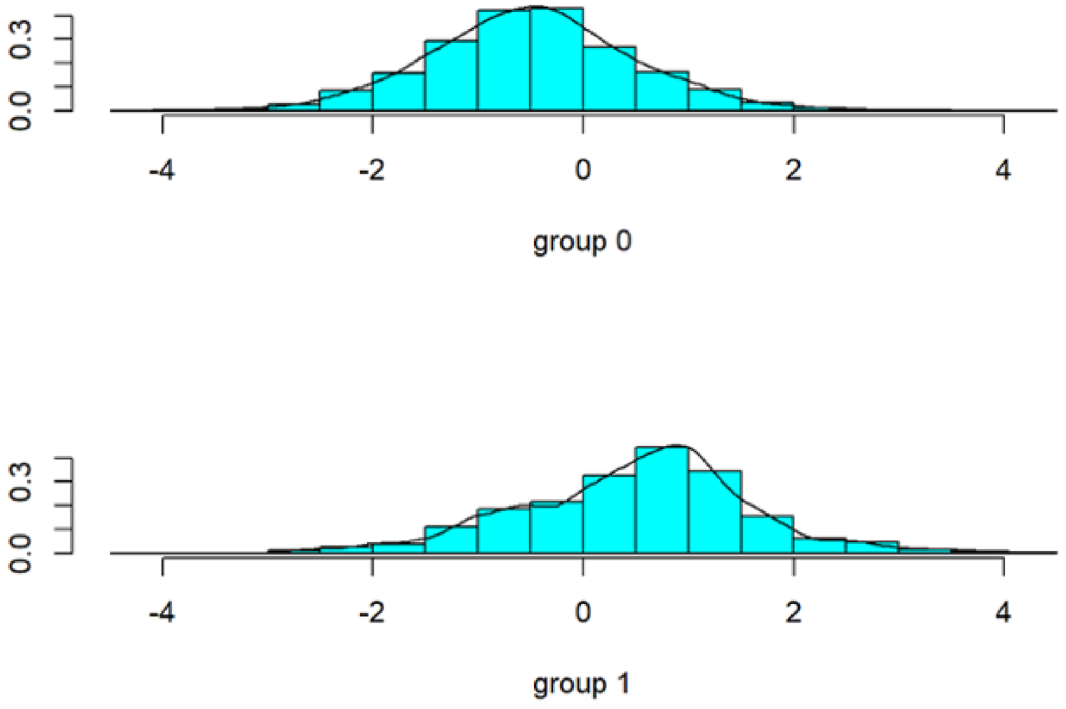


The coefficients also give a similar pattern and throw light on which X variables contribute most to group separation.

Step 6: Visualizing the Groups

Now, let's visualize what our non-churning customers look like compared to our customers who will most likely churn. Remember, 0 indicates customers who will not churn and 1 indicates customers who will.

```
plot(sublda, dimen = 1, type = "b")
```



The groups created by discriminant analysis can be seen in the graphs, and are in sync with the Wilks lambda value of 0.89 that we got from our MANOVA test. These graphs are a good indicator that although the model is significant, our two groups are not completely separated. There is some overlap.

Step 7: Make Predictions on the Test Set

Using our test set, let's predict whether a customer will churn or stay based on the X variables of the test set. We will apply the discriminant model that we built using the training set to make predictions about the test set. The objective of this is to measure how the model performs on a new set of data.

```
lda.pred=predict(sublda, newdata = testdata)

library(hmeasure)

class.lda=lda.pred$class
true.class<-testdata[,1]
lda.counts <- misclassCounts(class.lda,true.class)
lda.counts$conf.matrix
```

```
print(lda.counts$metrics,digits=3)
```

Step 8: Evaluate Model Performance Measures

The accuracy of the model is 1-Error rate = 1-0.1433 = 85.67%

Accuracy indicates how many correct predictions are made by the model. The model has a fairly good accuracy. However, since the dataset is unbalanced, accuracy alone may not be the sole indicator that the model is a robust model. Hence we will look at few other model performance measures.

Sensitivity also called the true positive rate is defined as the proportion of actual positives that are correctly identified by the model. The sensitivity of the model is 12.7% which is very low. For a churn prediction model, it is important that the model picks up positives as positives. It is important to make an accurate prediction of customers who will churn which is given by sensitivity.

We will now vary the threshold of the model from the default 50% to other values to decide on a optimum balance for sensitivity and specificity.

```
lda.pred$posterior[1:3,]
```

```
scores.lda <- lda.pred$posterior[,2]
all((scores.lda > 0.5)== (class.lda=="1"))
```

The model, by default, uses a 50% threshold to classify records as 0 or 1.

```
lda.counts.T03 <- misclassCounts(scores.lda>0.3,true.class)
lda.counts.T03$conf.matrix
```

```
lda.counts.T02 <- misclassCounts(scores.lda>0.2,true.class)
lda.counts.T02$conf.matrix
```

```
lda.counts.T017 <- misclassCounts(scores.lda>0.17,true.class)
lda.counts.T017$conf.matrix
```

```
lda.counts.T016 <- misclassCounts(scores.lda>0.16,true.class)
lda.counts.T016$conf.matrix
```

```
lda.counts.T015 <- misclassCounts(scores.lda>0.15,true.class)
lda.counts.T015$conf.matrix
```

```
lda.counts.T01 <- misclassCounts(scores.lda>0.1,true.class)
lda.counts.T01$conf.matrix
```

Now, let's compare the values of sensitivity and specificity for three threshold values.

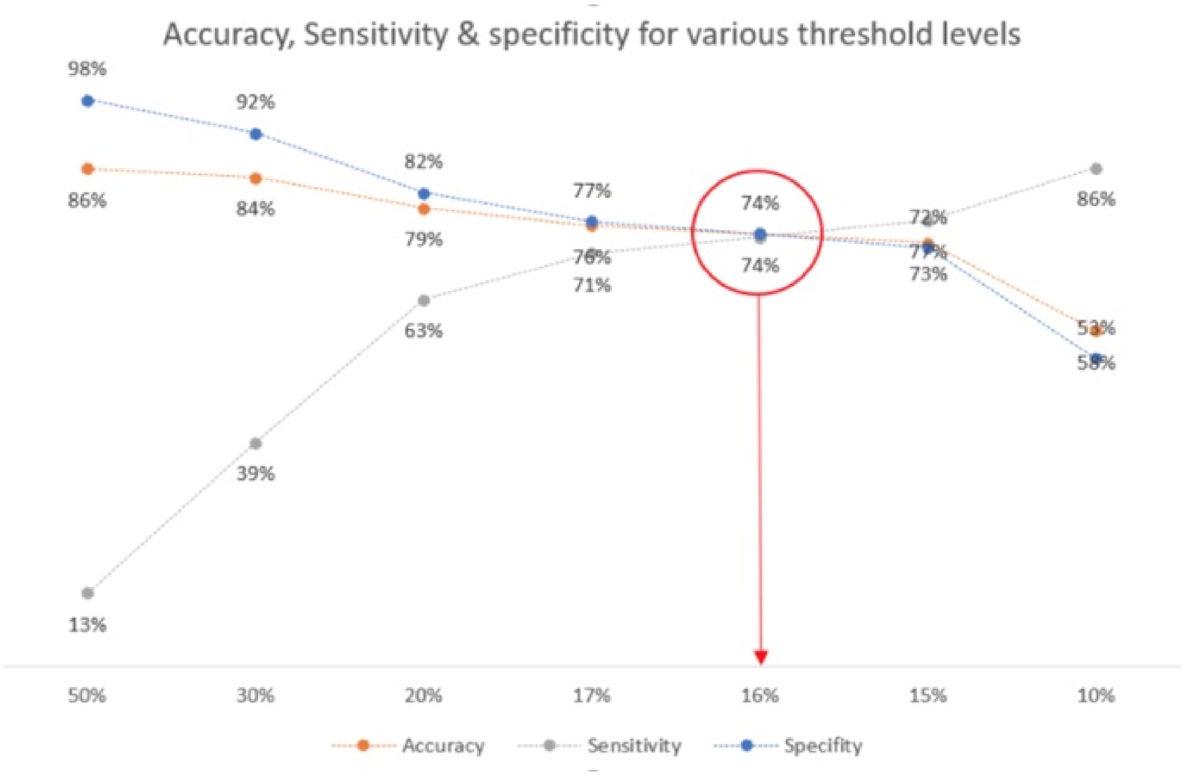
```
lda.counts.T02$metrics[c('ER', 'Sens','Spec')]
```

```
lda.counts.T017$metrics[c('ER', 'Sens','Spec')]
```

```
lda.counts.T016$metrics[c('ER', 'Sens','Spec')]
```

```
lda.counts.T015$metrics[c('ER', 'Sens','Spec')]
lda.counts.T01$metrics[c('ER', 'Sens','Spec')]
```

When we plot the accuracy, sensitivity, and specificity for various threshold values, the three lines intersect at a particular point. The threshold corresponding to this point indicates the optimum threshold for the model.



From the chart above, it is clear that the optimum threshold for the model is 16%. When we make predictions at this threshold, the accuracy, sensitivity, and specificity are 74%.

Insights From the Model and Business Recommendations

Based on the discriminant coefficients and the correl_ratio provided by the model, an increase in the below variables increases the probability of customer churn:

1. Number of customer service calls
2. Total day charge
3. Total evening charge
4. Total international charge
5. Total night charge

Additionally, an increase in number of voice mail messages decreases the probability of customer churn.

These insights from the discriminant model can help the business formulate strategies to reduce customer churn. Here's what I would recommend to the business based on what we've learned: First, customer issues should be resolved within the first or second call, as repeated calls to customer service causes customer churn. Second, there should be an organized escalation procedure for issues not resolved within two calls. Lastly, the provider should offer more attractive plans that reduce the cost of day, evening, and international calls based on usage.