**ITM 883**

**Business Analytics and Problem Solving**

**Project Report**

**Customer Churn Analysis**

A close up of text on a black background

Description automatically generated

**TEAM OUTLIERS:**

Karan Dalal - dalalkar@msu.edu

Neil Joshi - joshine2@msu.edu

Syed Kashif Mujtaba - kamoonpu@msu.edu

Vishal Agarwal - agarwa97@msu.edu

1. **OBJECTIVE**

In this project, we will analyze the historical customer data of a Telco organization to identify customers who are likely to churn and the factors that lead to a customer churning. We will be using logistic regression to classify whether a customer will churn or not.

1. **DATA OVERVIEW AND DESCRIPTION**

The dataset for this project has been obtained from Kaggle.com. The dataset has over 7000 rows and 20 predictor variables. Each row represents a customer and each column contains customer’s attributes.

|  |  |
| --- | --- |
| **COLUMN NAME** | **COLUMN DESCRIPTION** |
| Customer ID | A unique ID for each customer |
| Gender | Whether the customer is a male or a female |
| SeniorCitizen | Whether the customer is a senior citizen or not (1, 0) |
| Partner | Whether the customer has a partner or not (Yes, No) |
| Dependents | Whether the customer has dependents or not (Yes, No) |
| Tenure | Number of months the customer has stayed with the company |
| PhoneService | Whether the customer has a phone service or not (Yes, No) |
| MultipleLines | Whether the customer has multiple lines or not (Yes, No, No phone service) |
| InternetService | Customer’s internet service provider (DSL, Fiber optic, No) |
| OnlineSecurity | Whether the customer has online security or not (Yes, No, No internet service) |
| OnlineBackup | Whether the customer has online backup or not (Yes, No, No internet service) |
| DeviceProtection | Whether the customer has device protection or not (Yes, No, No internet service) |
| TechSupport | Whether the customer has tech support or not (Yes, No, No internet service) |
| StreamingTV | Whether the customer has streaming TV or not (Yes, No, No internet service) |
| StreamingMovies | Whether the customer has streaming movies or not (Yes, No, No internet service) |
| Contract | The contract term of the customer (Month-to-month, One year, Two year) |
| PaperlessBilling | Whether the customer has paperless billing or not (Yes, No) |
| PaymentMethod | The customer’s payment method (Electronic check, mailed check, Bank transfer, Credit card) |
| MonthlyCharges | The amount charged to the customer monthly |
| TotalCharges | The total amount charged to the customer (Combination of tenure and monthly charges) |
| Churn | Whether the customer churned or not (Yes or No) |

1. **METHODOLOGY**
2. **Exploratory Data Analysis**

In this phase, we performed exploratory data analysis to get insights using ggplot2 library.

A screenshot of a cell phone

Description automatically generated

*Fig. 1: The organization is successful in retaining majority of its customers*

A screenshot of a cell phone

Description automatically generated

*Fig. 2: Majority of the customers that churned were not senior citizens*

A screenshot of a cell phone

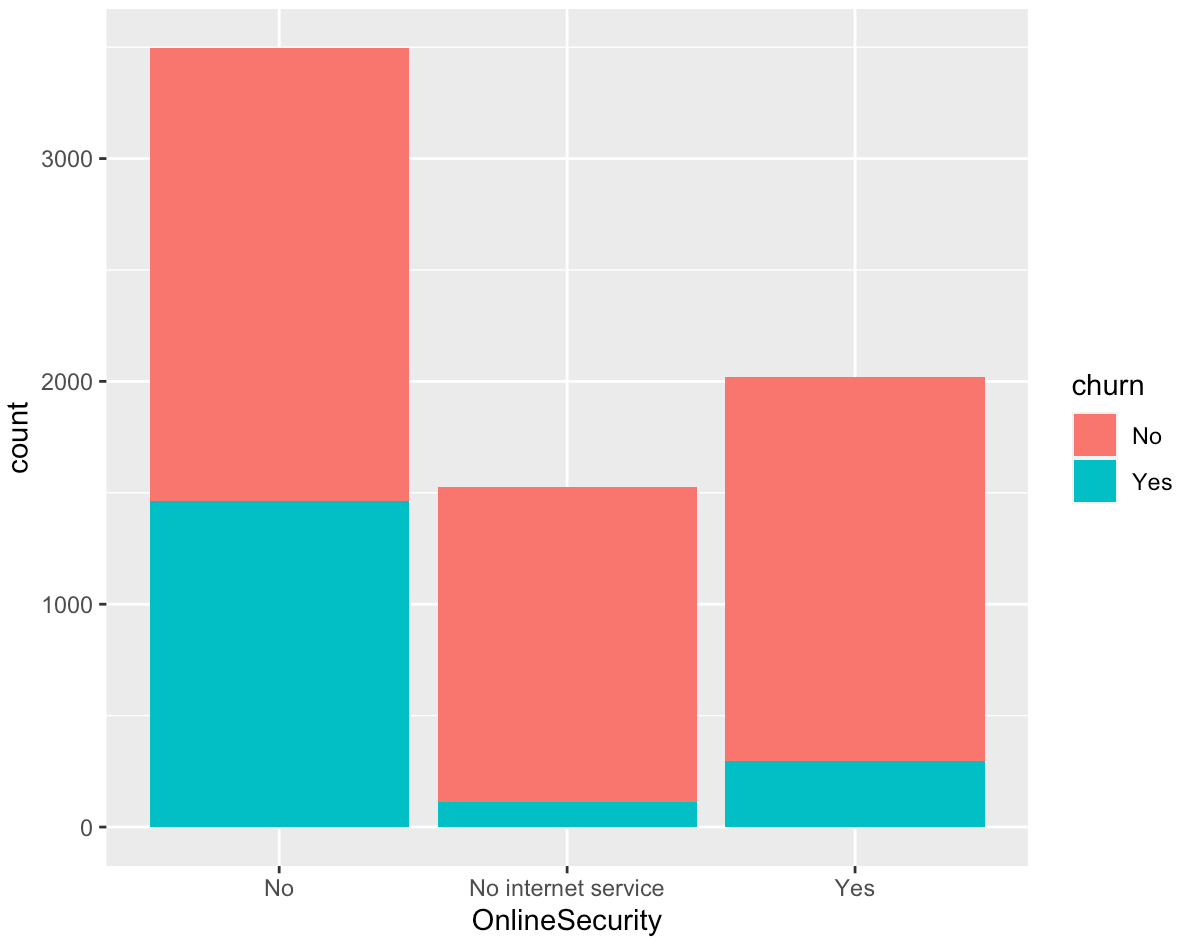
Description automatically generated

*Fig. 3: Majority of the customers that churned did not have dependents*

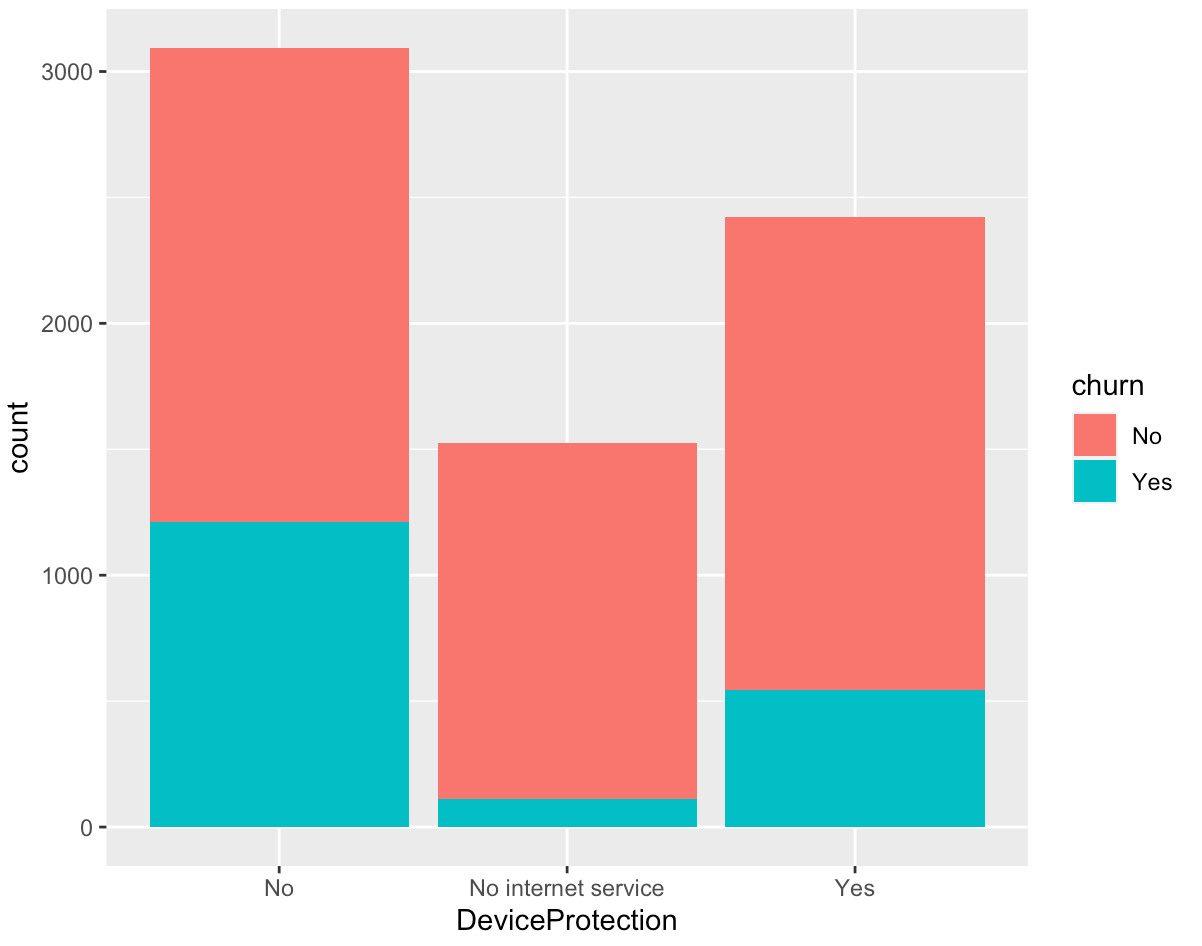
A screenshot of a cell phone

Description automatically generated

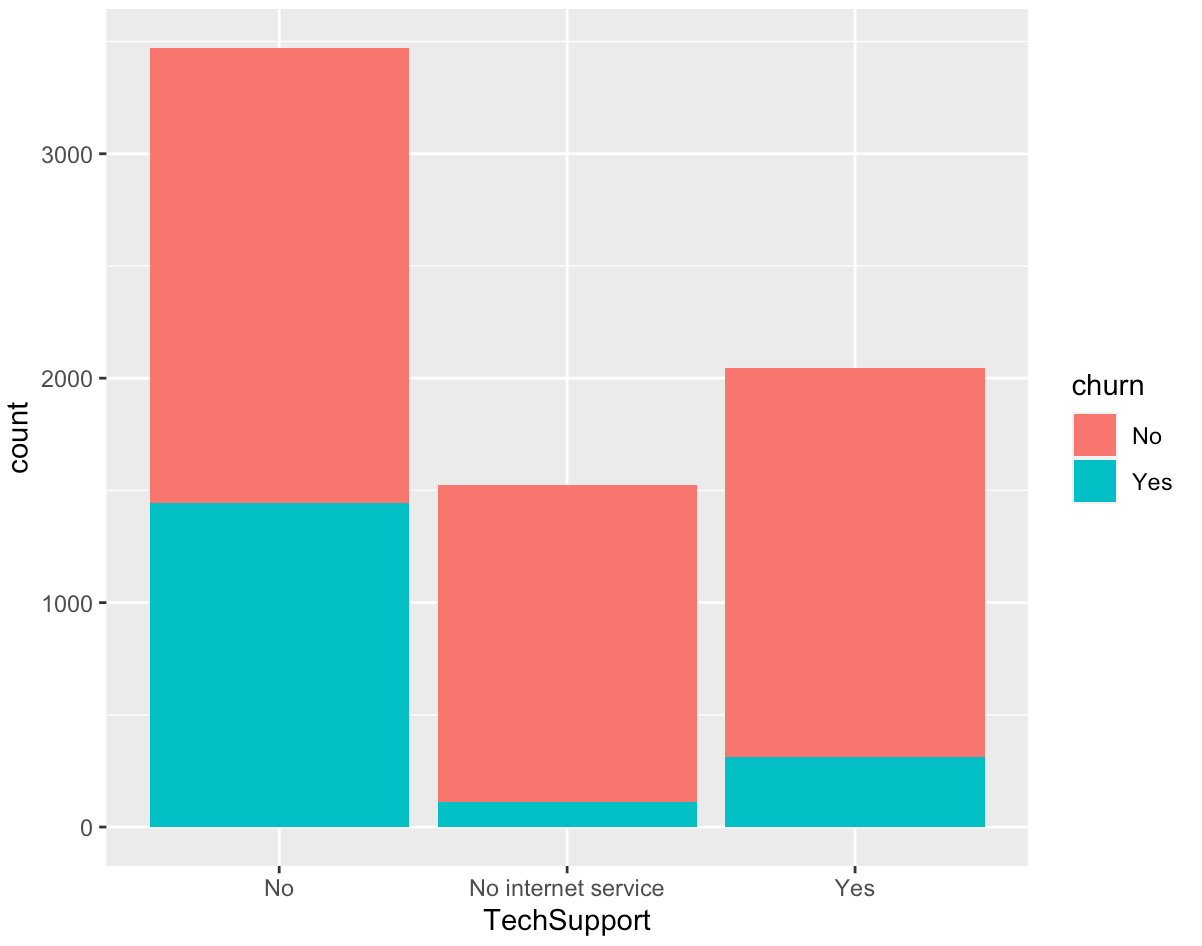
*Fig. 4: Majority of the customers that churned had phone service*



*Fig. 5: Majority of the customers that churned did not have Online Security*



*Fig. 6: Majority of the customers that churned did not have device protection*



*Fig. 7: Majority of the customers that churned did not have tech support*

A screenshot of a cell phone

Description automatically generated

*Fig. 8: Majority of the customers that churned had month-to-month contract*

A close up of a logo

Description automatically generated

*Fig. 9: Majority of the customers that churned had paperless billing*

A picture containing brick

Description automatically generated

*Fig. 10: Majority of the customers that churned paid through electronic check*

1. **Data Preprocessing**

We performed the following tasks during this phase of the project:

* Identified and imputed missing values
* Converted dependent variable (Churn) to binary format as follows:
* 0: The customer did not churn, i.e. he/she is still with the company
* 1: The customer churned, i.e. he/she is no longer a customer
* Converted numeric categorical variables into factors using as.factor()
* Removed multi-collinearity from the dataset

1. **Creating Logistic Regression Models**

For our analysis we created the following models:

* Model 1: Logistic Regression model using all the independent variable
* Model 2: Reduced Logistic Regression model using stepAIC()
* Model 3: Logistic Regression model with features selected using exhaustive search

For each of these models, the threshold value for classification has been selected using the ROC curve.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

*Fig. 11: ROC curve for Model 1, Model 2 and Model 3 respectively*

**Model 1:**

This model uses 23 independent variables to predict if a customer will churn or not. The best threshold value for this model is 0.286. The summary of the model is as follows:

A close up of text on a white background

Description automatically generated

*Fig. 12: Summary of Model 1*

**Model 2:**

This model was created by reducing Model 1 using the stepAIC() function. We have used ‘both’ direction to reduce this model. It has 18 independent variables. The best threshold value for this model is 0.27. The summary of the model is as follows:

A close up of a newspaper

Description automatically generated

*Fig. 13: Summary of Model 2*

**Model 3:**

This model has been created using the exhaustive search method. From the two graphs below, we can see that the best model should have 12 independent variables. The best threshold value for this model is 0.38. The summary of the model is as follows:

A close up of a map

Description automatically generated

A close up of a newspaper

Description automatically generated

*Fig. 14: (a) Adjusted R-squared and BIC graph, (b) Summary of Model 3*

1. **Evaluating the Models**

The confusion matrix for the models is as follows:

A picture containing table

Description automatically generated A picture containing table

Description automatically generated A picture containing table

Description automatically generated

*Fig. 15: Confusion Matrix for (a) Model 1, (b) Model 2, (c) Model 3*

We have calculated accuracy, precision, recall and error for all the models using confusion matrix.

A picture containing table

Description automatically generated A picture containing table

Description automatically generated A picture containing table

Description automatically generated

*Fig. 16: Accuracy, Precision, Recall and Error for (a) Model 1, (b) Model 2, (c) Model 3*

The specificity and sensitivity of the models are as follows:

A screenshot of a cell phone

Description automatically generated

*Fig. 17: Specificity and Sensitivity for (a) Model 1, (b) Model 2, (c) Model 3*

1. **Selecting the Best Model**

**Type 1 Error (FP):**

The model predicted that the customer will churn, but the customer did not actually churn.

**Type 2 Error (FN):**

The model predicted that the customer will not churn, but the customer actually churned.

We want our model to minimize Type 2 error. Since sensitivity can be calculated as:

*Sensitivity =*

We have selected Model 2 as our final model because it has the highest Sensitivity.

1. **Interpreting the Logistic Regression Model**
2. **Internet service with fiber optic**

Considering that all the other variables do not change, the odds that a customer churns increases by 6 times when the internet service is fiber optics rather than DSL

1. **Streaming Movies - Yes**

Considering that all the other variables do not change, the odds that a customer churns increases by 1.84 times if the customer streams movies against when customer does not streams movies

1. **Contract – Month to month**

Considering that all the other variables do not change, the odds that a customer churns increases by 2.05 times if the contract is month to month compared to yearly whereas the odds increases by 4.2 times when the contract is month to month compared to 2-year contract

1. **Payment Method (Electronic)**

Considering that all the other variables do not change, the odds that a customer churns increases by approx. 1.65 times if payment is made electronically vs bank transfer/credit card/mailed cheque

1. **Tenure**

As time the customer spent with the telco increases, the chances of the customer churning decreases

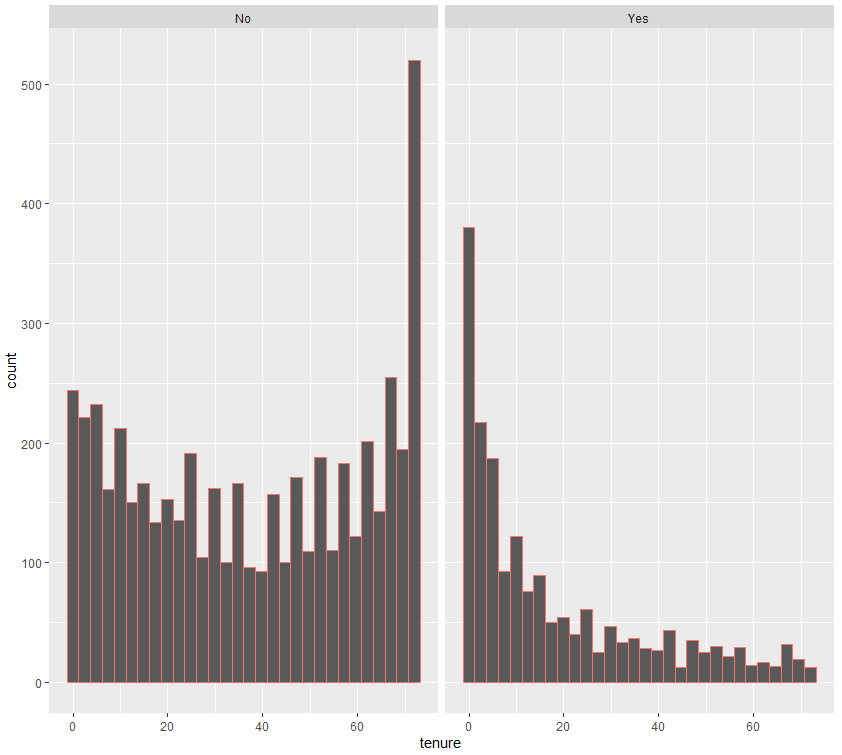
1. **RESULTS**

The way forward is a pyramid approach:

* Rectify the cracks and build a strong base
* Modification of the current model to suit the market dynamics
* Services/Promotions targeting reduction in churn

According to the model, the odds of the customer churning increases approximately 6 times if the customer uses Fiber Optic as Internet Service instead of DSL. To overcome this, the upper management of the Telco can come up with a promotion scheme that helps them retain these customers.

We also observed a significant increase in the odds of a customer churning if he/she is on a month-to-month contract rather than 1-year or 2-year contract. The upper management should build strategies to encourage customers to sign a 1-year or 2-year contract.



*Fig. 18: (a) Histogram of retained customers and tenure, (b) Histogram of churned customers and tenure*

From the graph shown in figure 18, we can see that most of the customers who churned had a tenure of less than 6 months. We would recommend the upper management to launch a loyalty bonus scheme, where customers get some concession on completing 1 year with the company.