Sri Lanka Institute of Information Technology

Fundamentals of Data Mining - IT3051

Group Number: G21

Student Number	Name
IT21176456	Gimmana M.R.M
IT21262104	Lakshani D. M. W. S
IT21261732	Perera W.A.S.K
IT21260988	Randeniya R.A.D.S.E

Contents

Background

Customer churn, also known as customer attrition or customer turnover, is a critical business metric that refers to the rate at which customers are going to stop doing business with a company or using its products or services during a specific period. It is an essential concept for businesses in various industries, including telecommunications, e-commerce, subscription services, and more. It directly impacts on a company's revenue and profitability because acquiring new customers often costs more than retaining existing ones.

Customers may have different behaviors, preferences, and reasons for canceling their subscriptions. By analyzing historical data and customer behavior, predictive models can be built to estimate the likelihood of a customer churning in the future. These models leverage historical data and various customer-related factors to make predictions about potential churners. It allows businesses to take steps to prevent churn before it happens. This information is invaluable for businesses seeking to take proactive measures to retain customers, improve customer satisfaction, reduce revenue loss, and optimize their marketing and customer service efforts. This data-driven approach is more efficient and cost-effective than reacting to customer losses after they occur, making it a valuable tool for customer retention and business growth.

Furthermore, as the business landscape becomes increasingly data-centric, the utilization of predictive models for churn analysis becomes a strategic imperative. These models not only identify customers at risk of churning but also unveil insights into the factors influencing their decisions. This enables businesses to tailor their interventions and offerings more precisely, enhancing the likelihood of customer retention. By proactively addressing customer churn, companies can foster stronger customer relationships, improve brand loyalty, and secure a sustainable position in today's competitive markets. Therefore, through this mini project, we aspire to provide a solution for reducing the rate of customer churn.

Scope of Work

The Scope of this mini project is to develop a Customer Churn Prediction system using Data Mining and Machine Learning techniques. The system will analyze the historical customer data to predict whether a customer is going to churn or not according to the customer's data.

The scope of this system has 5 layers,

1. User Interface Layer
2. Data Cleaning Layer
3. Data Mining Layer
4. Model Building and Analysis Layer
5. Data Visualization Layer

1.User Interface Layer

This layer is responsible for providing a user-friendly and easily interactive interface for users to input customer data and obtain churn predictions.

This layer focuses mostly on developing an environment that is user-friendly for the end user to engage with the backend analytics.

2.Data Cleaning Layer

This layer mostly handles the data's cleaning and preparation. Here, it recognizes incomplete, inaccurate, or irrelevant data sections and replaces, modifies, or deletes the unclean or coarse data after detecting and correcting corrupt or erroneous records. In this layer we are changing and mapping data from its raw form into a different format that is more suitable and useful for further analytical needs.

3.Data Mining Layer

This layer involves applying data mining techniques to the cleaned data. Various machine learning algorithms and statistical methods will be utilized to extract meaningful patterns, relationships, and insights from the customer data. This layer will involve tasks such as feature selection, dimensionality reduction, and exploratory data analysis.

4.Data Building and Analysis Layer

In this layer, predictive models will be built using the selected machine learning algorithms. The models will be trained on the customer data to learn patterns associated with customer churn. Model evaluation techniques such as cross-validation and performance metrics analysis will be employed to assess the accuracy and effectiveness of the models.

5. Data Visualization Layer

This layer focuses on visualizing the results and findings obtained from the data analysis and model building process. Data visualization techniques such as charts, graphs, and dashboards will be used to present the churn predictions, patterns, and insights in a visually appealing and easily interpretable manner.

Activities

- Data exploration and understanding of the data set.
    - Data preprocessing
    - Data splitting
    - Model selection
- Model Development
- Model Evaluation
- Make Predictions
- UI/UX Development

Approach

The approach to the project involves the following steps aimed at planning, executing, and completing the project successfully.

1. Dataset selection

The dataset we selected is the **Telecom Churn Prediction** dataset. Each row in the dataset represents a customer, each column contains the customer's attributes described in the column Metadata. The raw data contains 7043 rows (customers) and 21 columns (features).

Link to the dataset: https://www.kaggle.com/code/bandiatindra/telecom-churn-prediction/data

2. Data preprocessing

Data preprocessing involves preparing and cleaning raw data to make it suitable for training the model. It includes,

- Handling missing values
- Handling outliers
- Removing duplicate values
- Encoding categorical data

3. Feature selection

This step includes selecting the most suitable customer attributes to use as the inputs to the model for prediction. It is essential to select features that have the most impact on the target variable and remove irrelevant ones.

4. Data splitting

The preprocessed dataset will be divided into two parts as training set and test set. The training set will be included 80% and the testing set will be included 20% of the preprocessed dataset. The training set is used to train the model test set helps to evaluate the model's performance.

5. Model selection

Choosing the right machine-learning technique is crucial. Depending on the nature of the problem and the dataset characteristics, the technique we selected is **binary classification**.

6. Learning model

This involves training the selected machine learning model on the training dataset. The model learns patterns and relationships between the input features and target variables. The goal is to enable the model to make accurate predictions.

7. Applying model

Once trained, the model can be applied to the test dataset. The goal is to measure the accuracy of the model.

8. Analyzing

After obtaining the predictions, the model which gave the most accurate outputs will be selected to develop the system.

9. Building UI/UX

The method willing to use: Streamlet with Python.

10. Testing

This involves testing the accuracy of the model to identify and address issues.

Deliverables

In the first stage of our Telecom Churn Prediction project, we identified key stakeholders including our project team members and Telco Company. Team members are working on the project and the Telco Company is our stakeholder in this situation. We have defined a set of critical project deliverables to meet our needs and expectations.

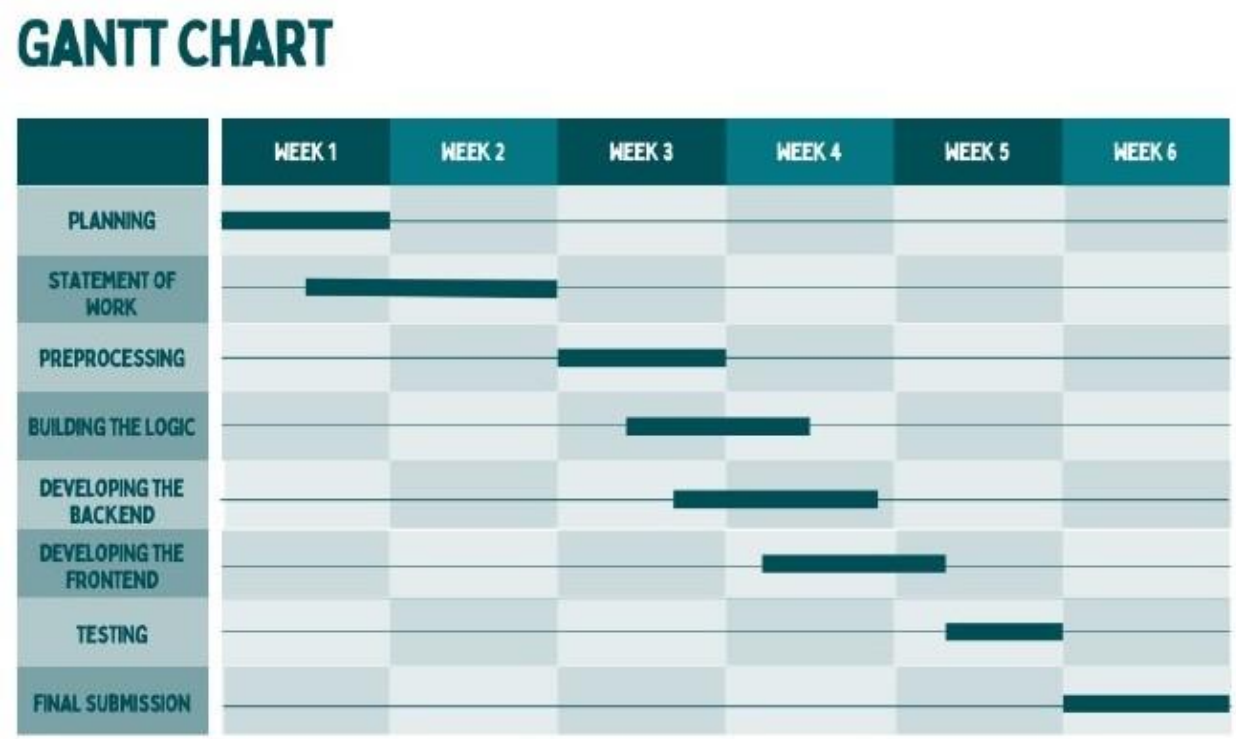
We have selected the Telecom Churn Prediction dataset, a comprehensive collection of customer data. Our aim is to predict a customer's likelihood of staying with the company using parameters such as phone service, internet service, gender, tenure, payment method, monthly charges, and total charges.

Our team meticulously cleaned the dataset, addressing missing values and outliers to ensure data quality. The dataset is properly formatted to facilitate analysis and model development. We use the machine learning model to predict if customers will remain with the company. The model undergoes training and evaluation, and its performance is evaluated using metrics such as precision, recall, accuracy, etc.

Our objective is to achieve a high accuracy forecasting model to efficiently manage project resources and make optimal use of team members' time as well as complete the project ahead of schedule.

To improve usability, we expose the model through a REST API, allowing users to interact with it and get predictions about a customer's likelihood to continue with the company.

Project Plan and Timeline



Assumptions

- The selected dataset will have relevant features and accurately labeled churn data.
- Adequate computational resources, including hardware and software, will be available to support model training and evaluation.
- The team members have the necessary programming skills and knowledge of Data Mining/Machine Learning techniques to successfully complete the project.
- The project will adhere to ethical considerations and data privacy regulations during data collection, storage, and analysis.

Project Team, Roles and Responsibilities

Name	Registration Number	Responsibilities
Gimmana M. R. M.	IT21176456	• Implement the model • Integrate • Data visualizing • Testing Data • Documentation • Implement user Interface
Randeniya R. A. D. S. E.	IT21260988	• Implement the model • Integrate • Data visualizing • Testing Data • Documentation • Implement user Interface
Lakshani D. M. W. S.	IT21262104	• Implement the model • Integrate • Data visualizing • Testing Data • Documentation • Implement user Interface
Perera W. A. S. K.	IT21261732	• Implement the model • Integrate • Data visualizing • Testing Data • Documentation • Implement user Interface