Phase-1 Submission Template

Student Name: Divya.v

Register Number: 511923205701

Institution: Priyadarshini Engineering College

Department: B.Tech-IT

Date of Submission: 5.05.2025

Topic: Predicting Customer churn using machine learning to uncover hidden patterns

GitHub Link:

1. Problem Statement

The insurance industry is a large financial service provider, with a billion-birr income. For it to grow, it is necessary to consider all the factors that lead to success and failure. In the business domain, machine learning increases the speed at which revenue can grow by using various models for various problems.

For this reason, the business model for machine learning that we have chosen is Customer Churn prediction in the case of Lion Insurance. Here, customers are switching from one insurance company to another for a variety of reasons, and as a result, the company is being adversely affected. While this goes on for years, it is important to identify loyal customers to manage the business and increase the number of customers and make a profit. Insurance companies must retain existing customers or attract new customers. But the chances of that happening are difficult to achieve. According to many researchers, various financial industries are operating in a highly competitive environment.

However, it is difficult to identify between the churner and the non-churners using traditional and manual work due to a large amount of data. Therefore, the process of model building is a complex task. The predictive models provide a way for the insurance company to attract new customers and retain the existing loyal customers. Customer management and decision-making can play a significant role in improving your organization's image and attracting new customers.

This gives an important strategy to keep customers and different strategic ways that must be considered by the insurance company.

Therefore, to solve this problem and address the gaps, it is necessary to know churners before they churn, so it seems very important to develop a model that predicts the future churners by applying machine learning techniques. This research also raises the following research questions: -

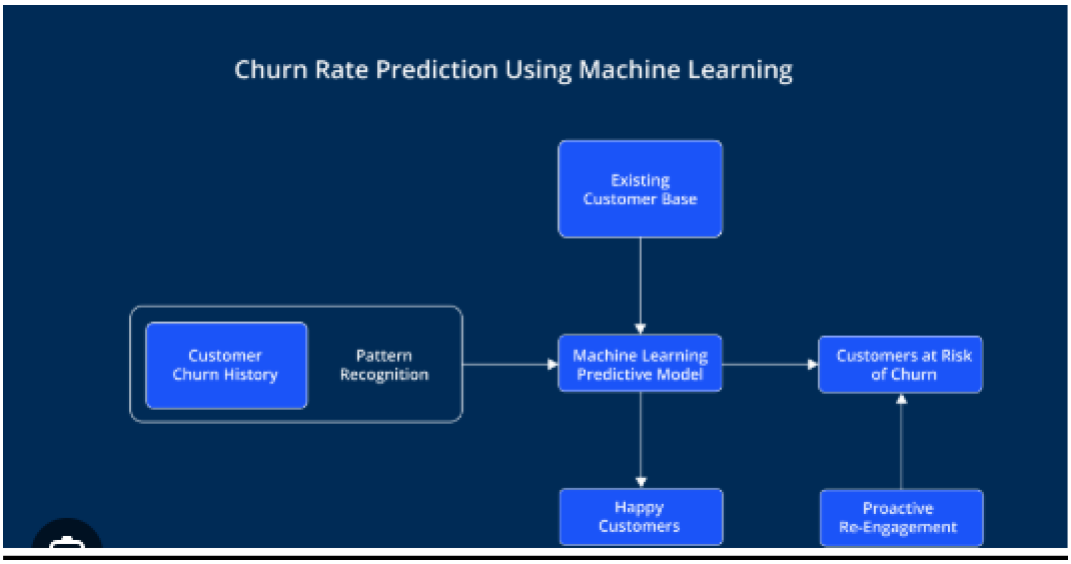
- Which customer characteristics are key to predicting customer churn behavior?
- How can unsupervised algorithms and classification algorithms be applied to customer churn prediction?
- Which churn prediction model generates the most accurate churn prediction results for the Lion Insurance Industry?

2. Objectives of the Project

The main objective of this project is to apply machine learning techniques to an existing motor insurance customer's data to obtain the best model which can predict churning customers. Specifically, this research aims to: -

- Reviewing concepts from literature and works related to customer churn prediction.
- Identify variables that are important to predict churn and non-churn.
- Perform data preparation tasks on the collected data for the correctness of the model building process.
- Develop the predictive models using different machine learning models.
- Evaluate the performance of the trained models to select the best model.
- Develop a prototype for the selected model.
- Provide concluding remarks and recommendations for further research works in this area.
- Identify limitations and future work on the proposed areas

3. Flowchart of the Project Workflow



4. Data Description

- **Dataset Name:** Telco Customer Churn
- **Source:** Kaggle
- **Type of data:** Structured tabular data
- **Records and Features:** The raw data contains 7043 rows (customers) and 21 columns (features).
- **Attributes Covered:** Customers who left within the last month – the column is called Churn

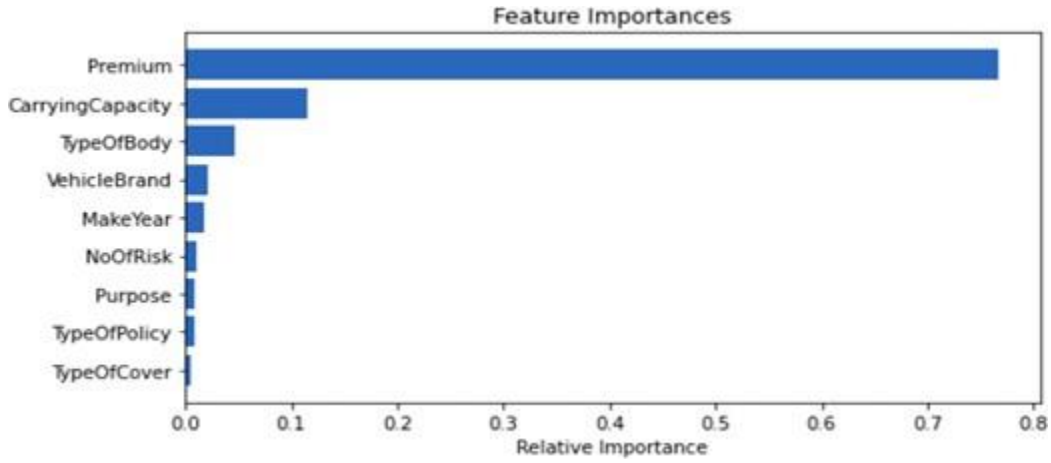
a. Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
b. Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
c. Demographic info about customers – gender, age range, and if they have partners and dependents

- **Dataset Link:** Telco Customer Churn

5. Data Preprocessing

- The Data Preprocessing section included cleaning missing values, Encoding, Normalizing the data (scaling), feature selection, cluster, sampling, and finally data splitting.
- In this study, there was one variable - "Number of Risks" missing values over 4%, and these were zero-filled. This is because this missing value indicates customers that are free from risk.
- Other categorical variables such as type of cover, vehicle brand, type of body, and car purpose missing values were filled by mode. Except for the number of risk variables other numerical variables' missing values were imputed by using the mean.
- For the sense of a common distribution, the dataset was scaled to a common scale by applying the Minmax normalization technique. The scaling technique was employed specifically on an individual column. Due to the value of the columns varied in min and max value.
- A correlation matrix has been used to find the relationship between independent and dependent variables and the issue of multicollinearity was detected and the most important feature has been identified in predicting the outcome.

- In addition to the correlation matrix, another feature selection method was used based on a previous study [28] to find the important predicting features. This method was also identified using sklearn.ensemble ExtraTreeClassifier function to identify the important predicting features.
- A graph of feature importance was plotted as shown below:



- As seen from the above graph also the most important feature in predicting the target was ‘Premium’ followed by ‘carrying capacity, ‘TypeOfBody’, and so on. K-means is one of the most popular and significant clustering methods exhibited by Mc. Queen in 1967. The data obtained from Lion Insurance is not labeled.
- To label, the data an unsupervised Kmeans algorithm was applied. When this algorithm is used, two classes were created.

- ❖ So to decide which data point is most likely to leave the company and which data points indicate the retaining one, the advice of an insurance professional was asked and labeled the dataset accordingly based on their comment and suggestions. For Experimentation modified Kmeans++ was used. From sklearn.cluster import kmeans_plusplus

6. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
    - Normality tests are used to determine if a dataset is normally distributed and to check how likely it is for a random variable in the dataset to be normally distributed.
    - Popular normality tests - D'Agostino's K^2, Shapiro-Wilk, Anderson-Darling .
    - There are three numerical features in this dataset
      - MonthlyCharges, Tenure, and TotalCharges.
    - Hypotheses -
    - H0: the sample has a Gaussian distribution.
    - H1: the sample does not have a Gaussian distribution.

- NB : we can not perform Shapiro-Wilk Test because sample size > 5000 and for this test p-value may not be accurate for N > 5000

- **Bivariate and Multivariate Analysis:**
  - In this section, I did an extensive statistical analysis with various hypotheses testing based on paired data types -
  - numerical and numerical data
  - numerical and ordinal data
  - ordinal and ordinal data
  - categorical and categorical data
  - General Hypotheses -
  - H0: the two samples are independent
  - H1: there is a dependency between the samples
- **Insights Summary**
  - Customer may churn depend upon the facilities produced by the particular industries.

7. Feature Engineering

a. machine learning can be used to predict customer churning in the B2B context.
b. churn prediction is one of two parts in customer churn management, and for future work, it would be interesting to investigate what features to use and how they impact churn prediction in the context of B2B.
c. Another aspect, which would be interesting to further investigate, is what other methods can be used for feature selection and sampling and how they would impact the result.

8. Model Building

- **Algorithm Used:**
    - Common classifiers, such as Naïve Bayes, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and Hidden Markov's Model are used in prediction problems
- **Model Selection Rationale**
    - XGBoost, which applies boosting and one algorithm.
    - Random Forest, which applies bagging
- **Train-Test-Split:**
    - 80% training,20%testing
- **Evaluation Metrics:**
    - evaluate their models using accuracy, precision, recall, and F-measure, which are all calculated from the confusion matrix.
    - case of imbalanced data, accuracy is not the most optimal metric used for evaluation

- This is why the data should be split into a training and a test set to be able to determine how well the model performs on examples not used during training.

9. Visualization of Results and Model Insights

❖ Feature Importance:

- is used in order to identify the most relevant features and is often used due to its performance enhancing properties
- Naïve Bayes assumes that each feature is independent, feature selection using feature importance was not performed.
- Advanced Model: recursive feature elimination (RFE)

❖ Model Comparison:

- machine learning can be used to predict customer churning within the domain of B2B;
- RandomUnderSampler for undersampling.

❖ Residual Plots

- Random Forest Classifier corresponds to 9589 classifications, which reflects a 99.75 percent accuracy.
- 98 percent non-churners, it is possible to obtain a 98 percent accuracy by predicting all examples as non-churners.

❖ User Testing

- After the unit testing, we must perform integration testing. The goal here is to see if modules can be integrated properly, the emphasis being on testing interfaces between modules. This testing activity can be considered as testing the design and hence the emphasis on testing module interactions.
- Here the entire software system is tested. The reference document for this process is the requirements document, and the goals to see if software meets its requirements.

- Acceptance Test is performed with realistic data of the client to demonstrate that the software is working satisfactorily. Testing here is focused on external behavior of the system; the internal logic of program is not emphasized

10. Tools and Technologies Used

- 1. Python
- a. Pandas(Python Library)
- b. Sklearn(Python Library)
- c. matplotlib(Python Library)
- d. imblearn(Python Library)
- e. Jupyter Notebook(Python Framework)

11. Team Members and Contributions

- **Data Cleaning: Tamil therndal B.I**
- **EDA: Divya**

Feature Engineering: Subulakshmi

- **Model Development: Vaishalini**
- **Documentation and Reporting: Sweatha**

.