Name :-Varsha V. Shinde

Batch no:-1838

Topic :-Blog/Article on Customer Churn Analysis Project.

Topics covered in this project:-

BLOG on Customer Churn Analysis

Introduction

Machine learning is a part of artificial intelligence and thereby a part of data science, which focuses on developing artefacts in the form of algorithms that learn from data and experience [27] [28]. By training algorithms on data, they can improve their decision-making and the accuracy of their predictions over time [27]. Machine learning is a frequently used approach considering automation of a variety of tasks and is categorized in different types of learning based on how the algorithms learn to become better in form of accuracy. These types of learning are usually categorized as the following: Supervised learning, unsupervised learning, semi supervised learning and reinforcement learning and based on the scope of the problem and the type of data used, one algorithm could be more suitable than the others

Te telecommunications sector has become one of the main industries in developed countries. Te technical progress and the increasing number of operators raised the level of competition [1]. Companies are working hard to survive in this competitive market depending on multiple strategies. Tree main strategies have been proposed to generate more revenues [2]:

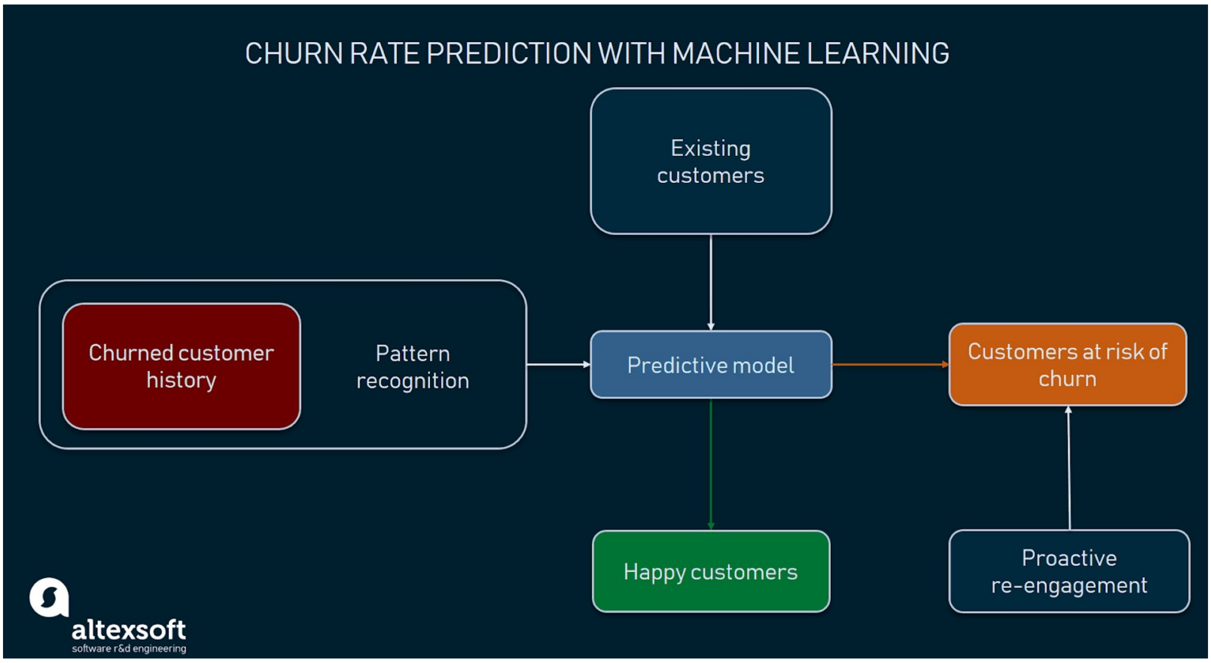
(1) Acquire new customers,

(2) Up sell the existing customers, and

(3) Increase the retention period of customers.

However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer costs much lower than acquiring a new one, in addition to being considered much easier than the up selling strategy .

To apply the third strategy, companies have to decrease the potential of customer’s churn, known as “the customer movement from one provider to another”.



CHURN RATE PREDICTION WITH MACHINE LEARNING

Existing customers

Churned customer history

Pattern recognition

Predictive model

Customers at risk of churn

Happy customers

Proactive re-engagement

PROBLEM DEFINITION

Customer churn or customer attrition is the phenomenon where customers of a business no longer purchase or interact with the business. A high churn means that higher number of customers no longer want to purchase goods and services from the business. Customer churn rate or customer attrition rate is the mathematical calculation of the percentage of customers who are not likely to make another purchase from a business.

Customer churn happens when customers decide to not continue purchasing products/services from an organization and end their association. It is an integral parameter for the organization since acquiring a new customer could cost almost 7 times more than retaining an existing customer. Customer churn can prove to be a roadblock for an exponentially growing organization and a retention strategy should be decided in order to avoid an increase in customer churn rates.

In order to capture the aforementioned problem, company should predict the customer's behaviour correctly. Customer churn management can be done in two ways:(1) Reactive & (2) Proactive. In the reactive approach, company waits for the cancellation request received from the customer, afterwards, company offers the attractive plans to the customer for the retention. In the proactive approach, the possibility of churn is predicted, accordingly the plans are offered to the customers. It's a binary classification problem where churners are separated from the non churners. In order to tackle this problem, machine learning has proved itself as a highly efficient technique, for forecasting information on the basis of previously captured data which includes linear regression, support vector machine, naïve bays, decision tree, random forest, etc.

In machine learning models, after pre-processing feature selection plays a significant role to improve the classification accuracy. A plenty of approaches were developed by researchers for feature selection that are useful to reduce the dimension, computation complexity & overfitting. In churn prediction, those feature are extracted from the given input vector which are useful for the prediction of churn. In this work, to tackle this problem we have used the following Machine Learning techniques:

- Logistic Regression,
- Naive Byes,
- Support Vector Machine.

Different Types of Churn

Customers cancel for various reasons and in multiple ways. You require a different set of action plans to tackle churn of each type. Here are some types of churn:

Voluntary Active Churn

These are customers that cancel your product or service. This type of Churn can occur due to various reasons, such as poor on boarding, poor customer service, or switching over to the competitor. This type of churn forms a large chunk of your lost revenue, and you should ocus most of your strategic initiatives on preventing it. Churn can occur due to various reasons.

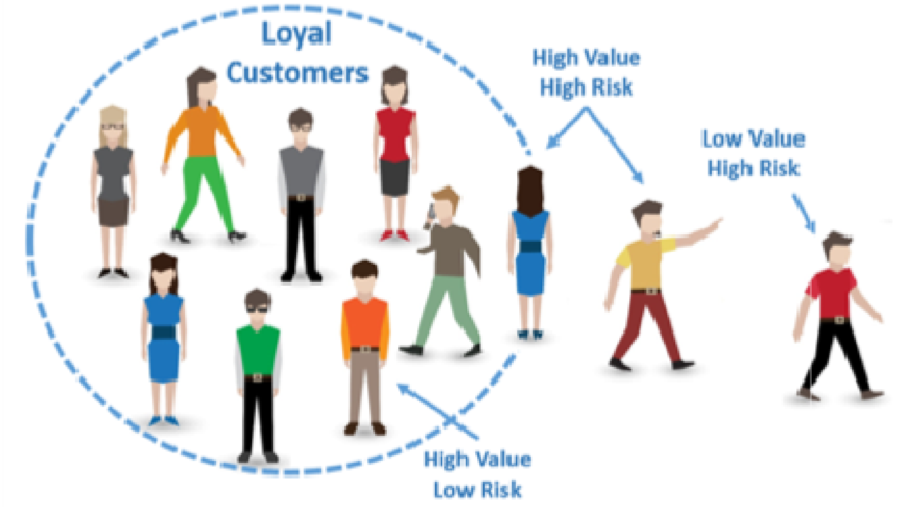
Involuntary Passive Churn

This type of churn is a leak in your revenue stream. Involuntary churn occurs when the :-

- When an expired card is used
- Hard declines happen when a card is reported lost or stolen.
- Soft declines occur when a credit card has maxed out its limit.
- Banks can decline the card (due to suspected fraudulent activity, frozen accounts, etc.)

DATA ANALYSIS:-

Objective is to reduce customer churn for Telecom service provider by identifying the potential churn customer before hands and take the proactive actions to make them stay. Customer churn is a loss to service provider as its difficult to get back the same customer due to bad marketing or other after sales services or user experience Major reasons why customer leaves or switch from the existing service provider Network issue / Billing Issue / best offers from other service providers / user experience not as per mark / Roaming charges & network while roaming / Data facility by service provider …etc…..We will build a logistic regression model to predict the customer churn of the firm based on the account information like Account Weeks, Contract Renewal, Data Plan, Data Usage, Overage Fee ,Roam Mins, and interpret the result.



EDA CONCLUDING REMARKS:-

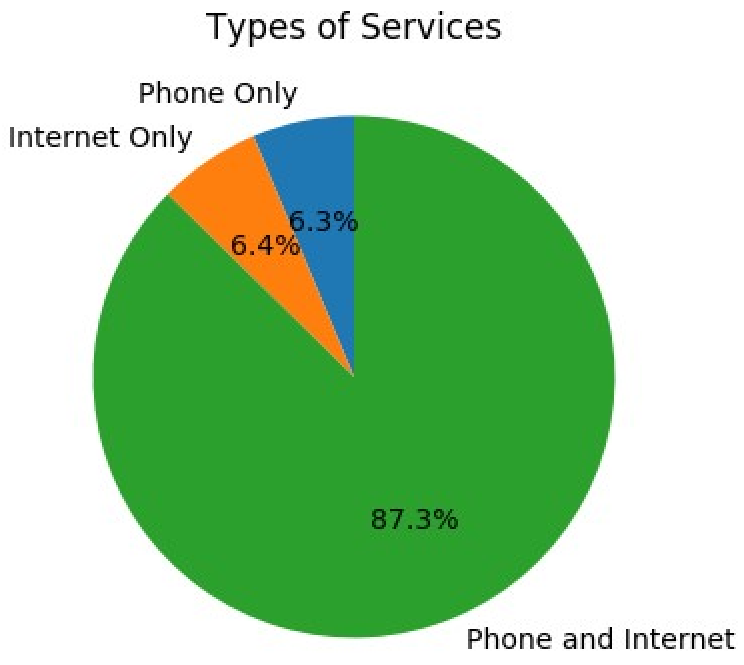
Exploratory data analysis (EDA) is an essential part of the data science or the machine learning pipeline. In order to create a robust and valuable product using the data, you need to explore the data, understand the relations among variables, and the underlying structure of the data.

EDA is **applied to investigate the data and summarize the key insights**. It will give you the basic understanding of your data, it's distribution, null values and much more. You can either explore data using graphs or through some python functions.

EDA is a phenomenon under data analysis used for gaining a better understanding of data aspect like:-

- Main Features of data
- Variable and relationship that hold between them
- Identifying which variables are important for our problem

Exploratory Data Analysis, or EDA, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset.



PRE-PROCESSING PIPELINE:-

Historical data that was selected for solving the problem must be transformed into a format suitable for machine learning. Since model performance and therefore the quality of received insights depend on the quality of data, the primary aim is to make sure all data points are presented using the same logic, and the overall dataset is free of inconsistencies. Previously we wrote an article about basic technique for dataset preparation.
The pre-processing of data involves 3 steps namely:-

1] Data Cleaning
2] Feature Selection
3] Data Transformation

Data transformation comprises of two explanatory variables which can be transformed from binomial form into binary form to be much application for the chosen models.
The data cleaning step involves missing data imputation or handling. some of the chosen algorithms cannot manage missing data that is why missing value can be transformed by median, mean or zero. However, the replacement of missing data by computed value statically is a better choice. The used set of data involve in certain numerical variables and two categorical variables.

BUILDING MACHINE LEARNING MODELS:-

The main goal of this project stage is to develop a churn prediction model. Specialists usually train numerous models, tune, evaluate, and test them to define the one that detects potential churners with the desired level of accuracy on training data.

Classic machine learning models are commonly used for predicting customer attrition, for example, logistic regression, decision trees, random forest, and others. Alex Bekker from Science Soft suggests using Random Forest as a baseline model, then *"the performance of such models as XGBoost, LightGBM, or Cat Boost can be assessed."* Data scientists generally use a baseline model's performance as a metric to compare the prediction accuracy of more complex algorithms.

Logistic regression is an algorithm used for binary classification problems. It predicts the likelihood of an event by measuring the relationship between a dependent variable and one or more independent variables (features). More specifically, logistic regression will predict the possibility of an instance (data point) belonging to the default category.

A **decision tree** is a type of supervised learning algorithm (with a predefined target variable.) While mostly used in classification tasks, it can handle numeric data as well. This algorithm splits a data sample into two or more homogeneous sets based on the most significant differentiator in input variables to make a prediction. With each split, a part of a tree is being generated. As a result, a tree with decision nodes and leaf nodes (which are decisions or classifications) is developed. A tree starts from a root node – the best predictor.

A **Random forest** is a type of an ensemble learning method that uses numerous decision trees to achieve higher prediction accuracy and model stability. This method deals with both regression and classification tasks. Every tree classifies a data instance (or votes for its class) based on attributes, and the forest chooses the classification that received the most votes. In the case of regression tasks, the average of different trees' decisions is taken.

CONCLUDING REMARKS:-

Churn rate is a health indicator for subscription-based companies. The ability to identify customers that aren’t happy with provided solutions allows businesses to learn about product or pricing plan weak points, operation issues, as well as customer preferences and expectations to proactively reduce reasons for churn.

It’s important to define data sources and observation period to have a full picture of the history of customer interaction. Selection of the most significant features for a model would influence its predictive performance: The more qualitative the dataset, the more precise forecasts are.

Companies with a large customer base and numerous offerings would benefit from customer segmentation. The number and choice of ML models may also depend on segmentation results. Data scientists also need to monitor deployed models, and revise and adapt features to maintain the desired level of prediction accuracy.

Customer churn analysis allows minimizing acquisition costs and increasing marketing efficiency, preparing a solid base for future marketing analysis and campaigns. Customer churn analysis opens new opportunities for cross-selling and up selling and serves as one of the starting points for customer-driven product development, keeping customers engaged and loyal over time.