BLOG/Articles Details

Customer Churn Prediction Using Machine Learning:

What is Customer Churn or Customer Attrition?

Customer churn is when a company's customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals

Customer retention refers to the ability of a company or product to retain its customers over some specified period. High customer retention means customers of the product or business tend to return to, continue to buy or in some other way not defect to another product or business, or to non-use entirely.

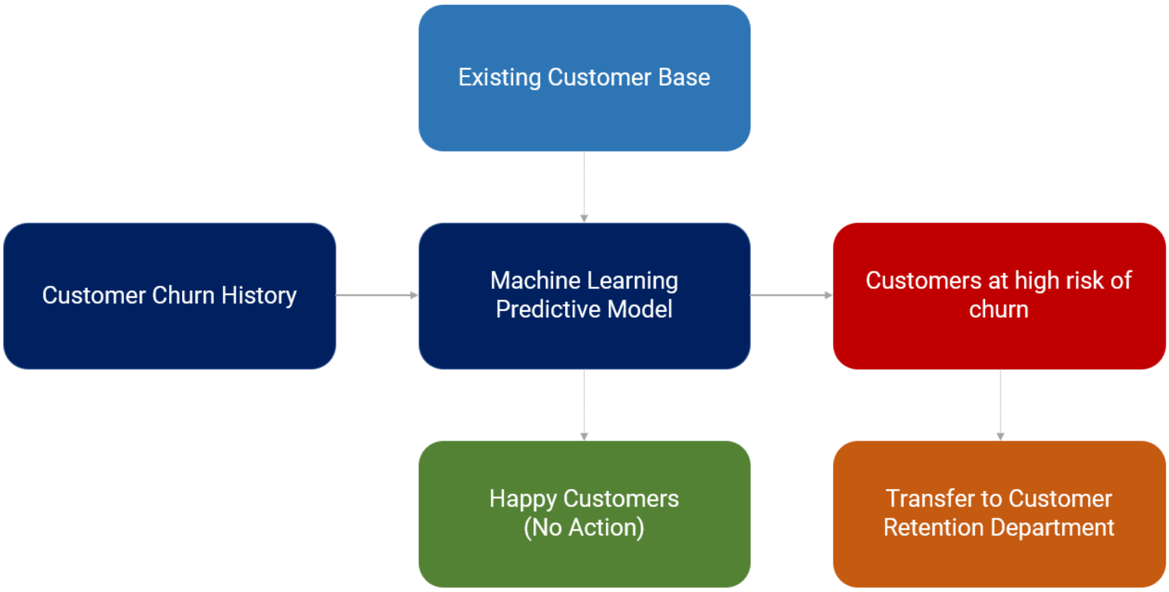
Customer retention starts with the first contact an organization has with a customer and continues throughout the entire lifetime of a relationship and successful retention efforts take this entire lifecycle into account.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base.

A company's ability to attract and retain new customers is related not only to its product or services, but also to the way it services its existing customers, the value the customers actually perceive as a result of utilizing the solutions, and the reputation it creates within and across the marketplace.

Successful customer retention involves more than giving the customer what they expect. Generating loyal advocates of the brand might mean exceeding customer expectations. Creating customer loyalty puts 'customer value rather than maximizing profits and shareholder value at the center of business strategy'. The key differentiation in a competitive environment is often the delivery of a consistently high standard of customer serviceFurthermore, in the emerging world Customer Successs, retention is a major objective.

Customer retention has a direct impact on profitability. Research by John Fleming and Jim Asplund indicates that engaged customers generate 1.7 times more revenue than normal customers, while having engaged employees and engaged customers return a revenue gain of 3.4 times the norm.



1. Customer Churn Analysis

Predicting customer churn with machine learning

While doing machine learning tasks,as a data science specialists we first need data to work with. Depending on the goal, researchers define what data they must collect. Then understanding a problem and final goal. Next, selected data is prepared, preprocessed, and transformed in a form suitable for building machine learning models. Finding the right methods to training machines, fine-tuning the models, and selecting the best performers is another significant part of the work. Once a model that makes predictions with the highest accuracy is chosen, it can be put into production.

Predicting customer churn with machine learning involves:

1. Problem Definition.
2. Data Analysis.
3. EDA Concluding Remark.
4. Pre-Processing Pipeline.
5. Building Machine Learning Models.
6. Conclusion.

1. Understanding a problem and a final goal
It's important to understand what insights one needs to get from the analysis. In short, you must decide what question to ask and consequently what type of machine learning problem to

solve: classification or regression. So when I analysed the data ,identified it as classification problem.As per my analysis I identified the data type and found that the target variable is an example of Binary Classification.

There should be a problem statement for every tasks .It describes about the task we have to predict.

Problem Statement:

Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

OBJECTIVE:

Examining customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

Identifying Type:

Target Variable is classic example of Binary Classification. Hence Logistic Regression algorithm best suits. However we will apply classification algorithms such as DecisionTree Classifier.

Dataset preparation and preprocessing

Data is the foundation for any machine learning project. The second stage of project implementation is complex and involves data collection, selection, preprocessing, and transformation.The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.The type of data depends on what you want to predict.

Data preprocessing

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Data formatting. The importance of data formatting grows when data is acquired from various sources by different people. The first task for a data scientist is to standardize record formats. A specialist checks whether variables representing each attribute are recorded in the same wayThe principle of data consistency also applies to attributes represented by numeric ranges.

Data cleaning. This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible. This stage also includes removing incomplete and useless data objects.

2.Exploratory Data Analysis and its Conclusion

Under this I performed some basic exploratory data analysis to get a better understanding of what is in our data such as:

- How much data we have
- If there are any missing values
- What data type each column is
- The distribution of data in each column

I could also take this opportunity to visualize data by plotting some charts to help us get an idea of what variables / features will prove useful. For example, if we were thinking of doing some regression analysis, scatter charts could give us a visual indication of correlation between features.Data visualization is large amount of information represented in graphic form is easier to understand and analyze. The visualization od datadepicts how the number of service calls and the use of international plans correlate with churn.

The pandas library has plenty of built in functions to help us quickly understand summary information about our dataset.I loaded data and converted into DataFrame. Below we use the shape() method to check how many rows are in our dataset and the describe() method to confirm whether or not our columns have missing values.

By using this df.columns code I checked all the columns available in the data set of customer churn analysis task.

```
print(df.dtypes)
```

```
print(df.info())
```

By using above these codes we can check data types of columns and we will get all information regarding the data we used.

3. Pre-Processing Pipeline

While analyzing data we had a huge dataset with mixed variables both numerical and categorical. So evaluated columns further. However columns like customerID is irrelevant, hence its better delete the column.

Checked missing values using the below code.

```
df.isnull().sum()
```

During this **exploratory data analysis** we got 11 missing values in our target variable Total Charges.Then did data viusualizations using various plots/grphs like countplot,pairplot,heat map etc. For Modelling we need to convert our target variable ‘Total Charges’ to binary digits.So used Label Encoder to convert it into binary. Now Target variable is converted into binary.

After cleaning and inspecting our data we might come to the conclusion that certain columns are not going to be useful for prediction. In this example we will not be using the phone-number of the client or geographical information about the client because our assumption is that this shouldn't affect churn.Column like customer Id also irrevalent so deleted that column too.

5.Building the Machine Learning Model and Testing

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 ScienceSoft suggests using Random Forest as a baseline model, then *"the performance of such models as XGBoost, LightGBM, or CatBoost can be assessed."* Data scientists generally use a baseline model's performance as a metric to compare the prediction accuracy of more complex algorithms.

At this point we can construct our model. The first thing to do is split our dataset into training and test sets. Given the ease of setting up a basic model, a common approach is to initialise and train a variety of different models and pick the most performance one as a starting point.

Here, we will first split our dataset into a training and validation set, so that we can train the model on the training set and evaluate its performance on the validation set. So that splitted dat into input and target.

We have split the data using the train_test_split function from the sklearn library keeping the test_size as 0.22 which means 22 percent of the total dataset will be kept aside for the validation set.

Once we have obtained our split we can use the RandomForestClassifier() , DecisionTreeClassifier(),AdaBoostClassifier(),LogisticRegression(),GradientBoostingClassifier() from the sklearn library as our models. After our model is trained, let's check its performance on both the training and cross validation set. We initialised our models, fit it to our dataset using the fit() method, then simply make our predictions using the predict() method. Then save the model.

Regression: Customer churn prediction can be formulated as a regression task. Regression analysis is a statistical technique to estimate the relationship between a target variable and other data values that influence the target variable, expressed in continuous values. If that's too hard – the result of regression is always some number, while classification always suggests a category. In addition, regression analysis allows for estimating how many different variables in data influence a target variable. With regression, businesses can forecast in what period of time a specific customer is likely to churn or receive some probability estimate of churn per customer.

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.

Decision tree basic structure.
Prediction results of decision trees can be easily interpreted and visualized. Even people without an analytical or data science background can understand how a certain output appeared. Compared to other algorithms, decision trees require less data preparation, which is also an advantage. However, they may be unstable if any small changes were made in data. In other words, variations in data may lead to radically different trees being generated. To address this issue, data scientists use decision trees in a group (AKA ensemble) that we'll talk about next.

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.

That's how Random Forest makes predictions
XGBoost is the implementation of the gradient boosted tree algorithms that's commonly used for classification and regression problems. Gradient boosting is an algorithm consisting of a group of weaker models (trees), which sums up their estimates to predict a target variable with more accuracy.

Conclusion:

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.

If we display the results we can see we have a list of booleans (0's and 1's) representing whether or not our model thinks a customer has churned or not. Now we can compare this to whether they actually churned to evaluate our model. We could also compute the actual probabilities of a customer churning using predict_proba() rather than just simple yes / no. We could then use these probabilities as a threshold for driving business decisions around which customers we need to target for retention, and how strong an incentive we need to offer them.

We can achieve the comparison mentioned above by using the .score() method, and displaying that we can see that we have achieved an accuracy of over 80%, which is not bad for our first attempt.

In our Telecom Customer Churn Analysis ,since GradientBoostingClassifier is performing best model among all other models which used, so saved it as final model. We are saving the model in pickle format and storing it as classifier.pkl. This will store the trained model and we will use this while deploying the model.

FEMINA ANEESH

2. Loan Application Status Prediction

The project that I have picked for this particular blog is automating the loan eligibility process. The task is to predict whether the loan will be approved or not based on the details provided by customers. Here is the **problem statement** for this project:

Problem Statement: Dataset includes details of applicants who have applied for loan. It includes details like credit history, loan amount, their income, dependents etc. To build a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.
Independent Variables:

Loan_ID

Gender

Married

Dependents

Education

Self_Employed

ApplicantIncome

CoapplicantIncome

Loan_Amount

Loan_Amount_Term

Credit History

Property_Area

Dependent Variable (Target Variable):

Loan_Status

Based on the details provided by customers, we have to create a model that can decide where or not their loan should be approved. This completes the problem definition part of the first stage of the machine learning lifecycle. After defining the problem statement we have to identify the type of the task.

IDENTIFYING TYPE : Target variable is categorical contains variable 'Y' or 'N', This is a classic case of Binary Classification and hence Logistic Regression algorithm best suits.However we will apply classification algorithms such as DecisionTree Classifier, RandomForest Classifier and SVC to verify best predicted model as per initial hypothesis.

Next, we need to **load the data and convert it into dataframe**.

We have some variables related to the loan, like the loan ID, which is the unique ID for each customer, Loan Amount and Loan Amount Term, which tells us the amount of loan in thousands and the term of the loan in months respectively. Credit History represents whether a customer has any previous unclear debts or not. Apart from this, we have customer details as well, like their Gender, Marital Status, Educational qualification, income, and so on. Using these features, we will create a predictive model that will predict the target variable which is Loan Status representing whether the loan will be approved or not.

Next are the **Data exploration and pre-processing** phase. Here, we will explore the dataset and pre-process it. The common steps under this step are as follows:

- Univariate Analysis
- Bivariate Analysis
- Missing Value Treatment
- Outlier Treatment
- Feature Engineering

We explore the variables individually which is called the univariate analysis. Exploring the effect of one variable on the other, or exploring two variables at a time is the bivariate analysis. We also look for any missing values or outliers that might be present in the dataset and deal with them. And we might also create new features using the existing features which are referred to as feature engineering. Again, I will not focus much on these data exploration parts and will only do the necessary pre-processing.

After exploring and pre-processing the data, next comes the model building phase. Since it is a classification problem, we can use any of the classification models like the logistic regression, decision tree, random forest, etc. I have tried logistic regression, GridSearchCV ,SVC models for this problem and random forest produced the best results. So, I will use a SVC as the predictive model for this project.

I have check the size of the dataset,shape of the dataset,column names of the dataset, data types of the data and all the other information of the data by importing required libraries.From this analysis we have a dataset of 614 samples and 13 features/columns. Most of the features are categorical and some are numerical. Target variable (Loan_Status) is categorical. Certain features like Loan_id, Loan_Amount_Term seems irrelevant in my initial analysis.

EDA Exploratory Data Analysis

I have checked the missing values in the columns. There are 13 null values in Gender, 3 null values in married, 15 null values in Dependents, 32 null values in Self_employed, 22 null values in Loan MAOUNT, 14 NULL VALUES IN lOAN amount term and 50 null values in Credit History. Lets fill the numeric Nan values with respective mean value. Also the Categorical value with mode value. Initially I tried another approach by dropping the Nan values using dropna(), However then I noticed more than 100 rows were deleted while dropping Nan values and hence changed the approach.

Filled the missed values for numerical terms by mean and filled the missed values for categorical terms by mode.Thus all missing values are replaced.

Then converted target variable that is categorical to numerical.For this I have used Label Encoding Technique.Label Encoding using pandas DataFrame replace function.Then checked the values of encoded variable.

After that done visualization of data using countplot,violinplots ,pairplots etc. From the visualization we can say that¶

-> Majority of the applicant income lies between 0 to 20000 and only few applicant have more than 20000 income

-> Majority of the coapplicant income lies between 0 to 7000 and only few coapplicant have more than 7000 income

->Majority of loan amount lies between 0 to 200.

Then transformed dataset into numerical category using Label encoder and dropped unnecessary columns like Applicant Income and Co-applicant Income.Then done statistical analysis and correlation between the variables.Next step is to checking skewness and removing the outliers.

MODELLING

Here, we will first split our dataset into a training and validation set, so that we can train the model on the training set and evaluate its performance on the validation set.So that splitted dat into input and target.

We have split the data using the train_test_split function from the sklearn library keeping the test_size as 0.22 which means 22 percent of the total dataset will be kept aside for the validation set. Next, we will train the GridSearchCV model ,LogisticRegression and SVC using the training set. After our model is trained, let’s check its performance on both the training and cross validation set.

The SVC model is 81% accurate on the validation set. Let’s check the performance on the training set too:

```
Accuracy Score = 0.8888888888888888
The CV Score = 0.8141173003241968
roc_auc_score 0.8261029411764707
```

Performance on the training set is almost similar to that on the validation set. So, the model has generalized well. Finally, we will save this SVC trained model so that it can be used in the future to make predictions on new observations:

We are saving the model in pickle format and storing it as classifier.pkl. This will store the trained model and we will use this while deploying the model.Saved SVC is the best Model.