INNOMATICS
RESEARCH LABS

Machine Learning - Project Report Document

Student Name	Relangi Kavya Sri
Batch	AI Elite 18
Project Name	Customer Churn Prediction
Project Domain	Telecommunication
Type of Machine Learning	Supervised ML
Type of Problem	Classification
Project Methodology	CRISP-DM

Stages Involved	Understand Business Problem • Exploratory Data Analysis • Data Pre-Processing and Feature Engineering • Building the Model using ML Algorithms • Evaluate the Model • Productionize the Model

Stage 1: Business Understanding:

Churn prediction is like having a crystal ball for business. It helps them anticipate when customers might leave or stop using their products or services. By understanding why customers leave and identifying warning signs early on, businesses can take proactive steps to keep them happy and engaged. This could mean offering specials deals, improving product features, or providing better customer support. Ultimately, churn prediction is about keeping customers around and ensuring the long-term success of the business.

Business constraints:

Telecom companies face hurdles in using churn data effectively. Inconsistent or missing data can lead to inaccurate predictions. Privacy regulations and internal silos might restrict access to useful customer information. Delays in data processing can make it difficult to take timely action to save customers at risk of churning. Additionally, the cost of implementing and maintaining churn prediction systems needs to be weighed against the potential revenue saved by retaining customers.

Stage 2: Data Collection and Understanding:

a)Data Collection:

Data collect from the client.

b)Data Understanding:

The data collected consists of various features related to customer attributes and services subscribed to. These features include customerID (unique identifier), gender, SeniorCitizen status, Partner status, Dependents status, tenure (months with the company), PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract type, PaperlessBilling status, PaymentMethod, MonthlyCharges, TotalCharges, and Churn status (indicating whether the customer churned or not.

S.NO	FEATURE NAME	DATA TYPE
1.	customerID	Object
2.	gender	Object
3.	Senior Citizen	Int64
4.	Partner	Object
5.	Dependents	Object
6.	Tenure	Int64
7.	PhoneServices	Object
8.	MultipleLines	Object
9.	InternetService	Object
10.	OnlineSecurity	Object
11.	OnlineBackup	Object
12.	DeviceProtection	Object
13.	TechSupport	Object
14.	StreamingTV	Object
15.	StreamingMovies	Object
16.	Contract	Object
17.	PaperBilling	Object
18.	PaymentMethod	Object
19.	MonthlyCharges	float64
20.	TotalCharges	Object
21.	Churn	Object

Stage 3: Data Preparation

a) Exploratory Data Analysis:

S.NO	TYPE	FEATURE NAMES	OBSERVATION
1.	Missing Values	Total Charges	Data type is in object is converted into 'float64'. Later found 11 missing values.

Observations and proper reasoning:

Handling missing values by replacing them with the median is a common approach, especially when the feature is numeric and not heavily skewed. By using the median instead of the mean, we avoid potential distortion caused by outliers. This method ensures that the imputed values are representative of the central tendency of the data and helps maintain the integrity of the dataset for subsequent analysis or modelling.

S no	Type of Cleaning	Technique	Feature Name	Reason
1	Missing value	Imputing with mean	Total Charges	Because outliers will be affected when we use mean.
2	Encoding	One hot	Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod.	Categorical column applied OHE that help to convert categorical to numerical.

3	Scaling	Standard Scaling	StandaradScaler	To Rescale/common scale.

Stage 4: Model Building:

S No	Type of Problem	Algorithm Name
1.	Classification	Logistic Regression
2.	Classification	KNN Classifier
3.	Classification	Decision Tree Classifier
4.	Classification	Random Forest Classifier
5.	Classification	SVC

Logistic Regression: Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables.

KNN: The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

Decision Tree Classifier: A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

Random Forest Classifier: Random forest is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

SVC: The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future.

Stage 5: Model Training:

Metrics used for evaluation:

The evaluation metric used for all algorithms during model training was accuracy, which measures the proportion of correctly classified instances over the total number of instances.

S No	Algorithm Name	Metric used for Evaluation
1.	Logistic Regression	Accuracy
2.	KNNeighbor Classifier	Accuracy
3.	Decision Tree Classifier	Accuracy
4.	Random Forest Classifier	Accuracy
5.	SVC	Accuracy

Stage 5: Model Evaluation:

S.NO	Algorithm Name	Evaluation
1.	Logistic Regression	0.812606
2.	KNNeighbor Classifier	0.768313
3.	Decision Tree Classifier	0.716070
4.	Random Forest Classifier	0.797842
5.	SVC	0 . 804656

Challenges Faced:

1.) It took some time to find the missing values in Total Charges because the missing values are stored in format of Object. At least 11 values are found with empty string “ “.

Conclusion:

Based on the Accuracy metrics, the best model for the classification problem appears to be Logistic Regression model. Although all the other models had similar accuracy scores, Logistic Regression model had the highest Accuracy.

While Logistic Regression model had a good accuracy score of 0.8126. Therefore, based on the metrics evaluated Logistic Regression model appears to be the best model for this classification problem.