Machine Learning - Project Report Document

Student Name	Sada Vijay
Batch	AI Elite 18
Project Name	Churn Prediction
Project Domain	Telecommunications
Type of Machine Learning	Supervised ML
Type of Problem	Classification
Project Methodology	CRISP-DM
Stages Involved	• Data Collection and Understanding • Data Preparation • Model Building • Model Training • Model Evaluation

Business Understanding:

The telecommunications sector has become one of the main industries in developed countries. The technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies.

Three main strategies have been proposed to generate more revenues:

1. Acquire new customers
2. Upsell the existing customers
3. Increase the retention period of customers

Customer churn is a considerable concern in service sectors with highly competitive services. On the other hand, predicting the customers who are likely to leave the company will represent a potentially large additional revenue source if it is done in the early phase.

Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data.

Customer Churn is one of the most important and challenging problems for businesses such as Credit Card companies, cable service providers, SASS and telecommunication companies worldwide. Even though it is not the most fun to look at, customer churn metrics can help businesses improve customer retention.

Problem Statement: Using customer information such as Gender, Senior Citizen status, Partnership status, and Family dependents, etc. develop a model to predict whether a customer will churn or not.

Here are some potential business constraints:

1. Limited resource allocation.
2. Data availability and quality.
3. Model interpretability.
4. Regulatory compliance.
5. Balancing customer engagement.
6. Monitoring competitive landscape.
7. Time sensitivity in churn prediction.

Stage 1: Data Collection and Understanding

a) **Data Collection:** The data was provided to us by the client.

b) **Data Understanding:**

Here are the features and their descriptions:

1. Gender: The gender of the customer.
2. Senior Citizen: Whether the customer is a senior citizen or not.
3. Partner: Whether the customer has a partner or not.
4. Dependents: Whether the customer has dependents or not.
5. Tenure: The duration for which the customer has been with the service provider.
6. Phone Service: Whether the customer has signed up for phone service or not.
7. Multiple Lines: Whether the customer has signed up for multiple phone lines or not.
8. Internet Service: The type of internet service the customer has signed up for (e.g., DSL, Fiber Optic, None).
9. Online Security: Whether the customer has signed up for online security service or not.
10. Online Backup: Whether the customer has signed up for online backup service or not.
11. Device Protection: Whether the customer has signed up for device protection service or not.
12. Tech Support: Whether the customer has signed up for tech support service or not.
13. Streaming TV: Whether the customer has signed up for streaming TV service or not.
14. Streaming Movies: Whether the customer has signed up for streaming movies service or not.
15. Contract: The type of contract the customer has (e.g., Month-to-month, One year, Two years).
16. Paperless Billing: Whether the customer has opted for paperless billing or not.
17. Payment Method: The method of payment chosen by the customer (e.g., Electronic check, Mailed check, Bank transfer, Credit card).
18. Monthly Charges: The amount charged to the customer on a monthly basis.
19. Total Charges: The total amount charged to the customer.
20. customer ID: The individual identity of the customer.
21. Churn: Whether the customer left within the last month or not.

S No	Feature Name	Data Type
1	Gender	Object
2	Senior Citizen	Int64
3	Partner	Object
4	Dependents	Object
5	Tenure	Int64
6	Phone Service	Object
s7	Multiple Lines	Object
8	Internet Service	Object
9	Online Security	Object
10	Online Backup	Object
11	Device Protection	Object
12	Tech Support	Object
13	Streaming TV	Object
14	Streaming Movies	Object
15	Contract	Object
16	Paperless Billing	Object
17	Payment Method	Object
18	Monthly Charges	float64
19	Total Charges	Object
20	Customer ID	Object
21	Churn	Object

Stage 2: Data Preparation

a) Exploratory Data Analysis:

S No	Type	Feature Names	Observation
1	Missing Values	Total Charges	There are some empty strings present in this column.
2	Duplicates	NA	NA
3	Outliers	NA	NA

b) Data Cleaning/wrangling:

S no	Type of Cleaning	Technique	Feature Name	Reason
1	Missing value	Imputing with median	Total Charges	Replaced the missing values i.e. empty strings
2	Encoding	One hot	gender, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method	Used One Hot Encoding since the data in these categorical columns are nominal.
3	Scaling	Standard Scaling	Senior Citizen, tenure Monthly Charges, Total Charges	Used standardization to scale down all the columns into a similar scale ranging between 0 to 1 based on standard deviation.

Stage 3: Model Building:

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

1. **Logistic Regression:** Logistic regression is a statistical method used to predict the probability of an event happening, such as whether an email is spam or not. Unlike linear regression, it works well for situations where the outcome is binary (yes/no) instead of continuous.

2. **SVC:** A support vector classifier (SVM) excels at finding the best separation line between categories in your data. It prioritizes a wide margin between the classes, making it effective even for complex datasets.

3. **KNeigbors Classifier:** The K-Nearest Neighbors (KNN) classifier predicts a data point's class by analyzing the labels of its closest neighbors in the training data, making it simple to understand and effective for various classification tasks.

4. **Decision Tree Classifier**: A decision tree classifier is a machine learning method that uses a tree-like structure to classify data. It asks a series of questions about the data's features, branching out based on the answers, until it reaches a final leaf node that predicts the class.

5. **Random Forest Classifier:** Random Forest Classifier is a machine learning algorithm that combines multiple decision trees for stronger predictions. By training a "forest" of trees on random subsets of data, it reduces the risk of overfitting and improves overall accuracy.

Stage 4: Model Training:

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

Stage 5: Model Evaluation:

S No	Algorithm Name	Metric Score
1	KNN	0.768881
2	Logistic Regression	0.816014
3	SVC	0.809199
4	Random Forest Classifier	0.797842
5	Decision Tree Classifier	0.733674

Challenges Faced:

While identifying the missing values inside the total charges column, we found that it wasn’t an empty string but a single space character “ “ that was inside the data which made the column into object datatype.

Conclusion:

From the above Accuracy results we can observe that the Logistic Regression model has the highest accuracy when compared with the other models. We can see that the Logistic Regression model has the accuracy of 0.816. Therefore, we can say that the Logistic Regression appears to be the best model for classification task on our dataset based on the evaluation metrics.