REPORT

Predicting Customer Churn in a Telecommunications Company

Name - Soham Chakraborty
Roll_Number - CSE/20081/601
Email_Id- soham20601@iiitkalyani.ac.in
Year - 2020-2024

1.Data Collection And Preprocessing

1.1 Data Collection ➖

The dataset used for this analysis is sourced from kaggle.
Here is the link : https://www.kaggle.com/datasets/blastchar/telco-customer-churn

1.2 Data Preprocessing ➖

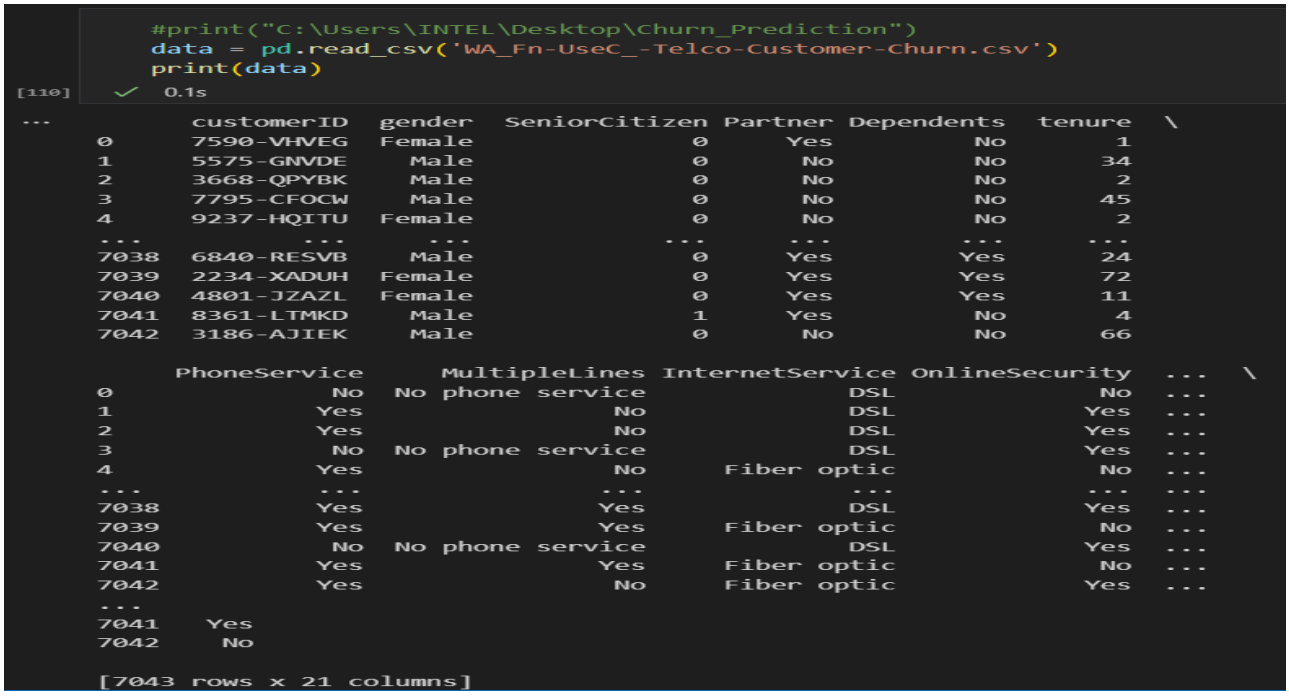
1. Loaded the dataset and observed its structure.

2. Handled missing values in the TotalCharges column by converting it to numeric and dropping rows with missing values.

3. Removed customer IDs from the dataset.

4. Converted the target variable Churn into binary numeric values (0 for 'No' and 1 for 'Yes').

5. Converted categorical variables into dummy variables using one-hot encoding.



```
    #print("C:\Users\INTEL\Desktop\Churn_Prediction")
    data = pd.read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv')
    print(data)
[110]  ✓ 0.1s
...          customerID  gender  SeniorCitizen Partner Dependents  tenure  \
0          7590-VHVEG  Female              0     Yes         No       1
1          5575-GNVDE    Male              0      No         No      34
2          3668-QPYBK    Male              0      No         No       2
3          7795-CFOCW    Male              0      No         No      45
4          9237-HQITU  Female              0      No         No       2
...               ...     ...            ...     ...        ...     ...
7038       6840-RESVB    Male              0     Yes        Yes      24
7039       2234-XADUH  Female              0     Yes        Yes      72
7040       4801-JZAZL  Female              0     Yes        Yes      11
7041       8361-LTMKD    Male              1     Yes         No       4
7042       3186-AJIEK    Male              0      No         No      66

     PhoneService     MultipleLines InternetService OnlineSecurity  ...  \
0              No  No phone service             DSL             No  ...
1             Yes                No             DSL            Yes  ...
2             Yes                No             DSL            Yes  ...
3              No  No phone service             DSL            Yes  ...
4             Yes                No     Fiber optic             No  ...
...           ...               ...             ...            ...  ...
7038          Yes               Yes             DSL            Yes  ...
7039          Yes               Yes     Fiber optic             No  ...
7040           No  No phone service             DSL            Yes  ...
7041          Yes               Yes     Fiber optic             No  ...
7042          Yes                No     Fiber optic            Yes  ...
...
7041    Yes
7042     No

[7043 rows x 21 columns]
```

2. Exploratory Data Analysis

2.1 Descriptive Statistic

Calculated and visualized descriptive statistics for numerical features.

```
pred.head()
```

[112] ✓ 0.0s Python

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSup
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	

5 rows × 21 columns

```
pred.columns.values
```

[113] ✓ 0.0s Python

```
array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

2.2 Visualizations

1. Explored the distribution of gender.
   Visualized the gender distribution with percentage annotations.
   Analyzed the distribution of senior citizens.

2. Presented the distribution as a pie chart.
   Explored customers with dependents and partners.

3. Visualized the percentage of customers with dependents and partners using a stacked bar chart.
   Examined the distribution of tenure.

4. Utilized a histogram to show the distribution.
   Investigated the distribution of contract types.

5. Displayed the count of customers based on contract types using a bar chart.
   Explored the distribution of tenure across different contract types.

6. Presented histograms for Month-to-Month, One Year, and Two Years contracts.
   Analyzed services such as phone service, multiple lines, internet service, and others.

7. Visualized the count of customers for each service type.
   Explored the relationship between monthly and total charges.

8. Created a scatter plot to visualize the correlation.
   Visualized the churn rate.
   Displayed the churn rate with percentage annotations.

.

3. Building the Churn Prediction Model

3.1 Model Selection ➖

Logistic Regression machine learning model is used to build the project.
And I got **80% accuracy** by using the **Logistic Regression Model**.

```
[136]  from sklearn.model_selection import train_test_split
       X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
       ✓ 0.0s

[137]  from sklearn.linear_model import LogisticRegression
       model = LogisticRegression()
       result = model.fit(X_train, y_train)
       ✓ 0.1s
```

```
[38]   from sklearn import metrics
       prediction_test = model.predict(X_test)
       # Print the prediction accuracy
       print (metrics.accuracy_score(y_test, prediction_test))
       ✓ 0.0s

       0.8075829383886256
```

.

3.2 Model Evaluation ➖

1. Imported the metrics module from scikit-learn.
2. Made predictions on the test set using the trained model.
3. Printed the accuracy score of the churn prediction model.

4. Conclusion and Recommendations

1. Summarized key findings from the analysis.
2. Provided insights into customer behavior and factors influencing churn.
3. Recommended areas for further investigation or actions to reduce churn.

5. Future Work

Feature Engineering:

Experiment with additional feature engineering techniques to create new meaningful features that might enhance the predictive power of the model. This could include interaction terms, polynomial features, or other transformations.

Hyperparameter Tuning:

Perform a more extensive hyperparameter tuning to optimize the logistic regression model. Grid search or random search can be used to explore a broader range of hyperparameter combinations.

Feature Selection:

Investigate feature selection methods to identify the most relevant features for the logistic regression model. This can help improve model interpretability and potentially reduce overfitting.

Addressing Imbalanced Data:

If the dataset is imbalanced (i.e., there is a significant class imbalance between churn and non-churn instances), explore techniques such as oversampling, undersampling, or the use of different class weights to handle imbalanced classes.

Ensemble Methods:

Explore the use of ensemble methods, such as bagging or boosting, to combine multiple logistic regression models. This may improve predictive performance by leveraging the strengths of different models.

Model Evaluation Metrics:

Consider using alternative evaluation metrics that are more relevant to the business context. For example, precision, recall, and F1-score may be more informative than accuracy in the context of churn prediction.

Time-Series Analysis:

If your dataset includes a time component, consider incorporating time-series analysis techniques. This could involve exploring trends and patterns in customer behavior over time to better understand the temporal aspects of churn.

Customer Segmentation:

Explore customer segmentation techniques to identify different groups of customers with distinct characteristics. Building separate models for different segments may lead to more accurate predictions.

Feedback Loop:

Implement a feedback loop to continuously update and improve the model based on new data. Regularly retrain the model to adapt to changing customer behavior patterns.