Sukanya

Capstone Project - 2

Telecom Churn Dataset (IBM Watson Analytics)

Sukanya

Customer Churn

- Customer churn plays a huge role in the performance of service companies.
- To manage churn it is necessary to understand relation between churn and the customer costs

Sukanya

Dataset Source

Telecom Churn Dataset (IBM Watson Analytics)

url:
https://www.kaggle.com/datasets/zagarsuren/telecom-churn-dataset-ibm-watson-analytics/data

Why Customer Retention

“Loyal customers, they don't just come back, they don't simply recommend you, they insist that their friends do business with you.” “If you are not taking care of your customers, your competitor will.” - Chip Bell

Sukanya

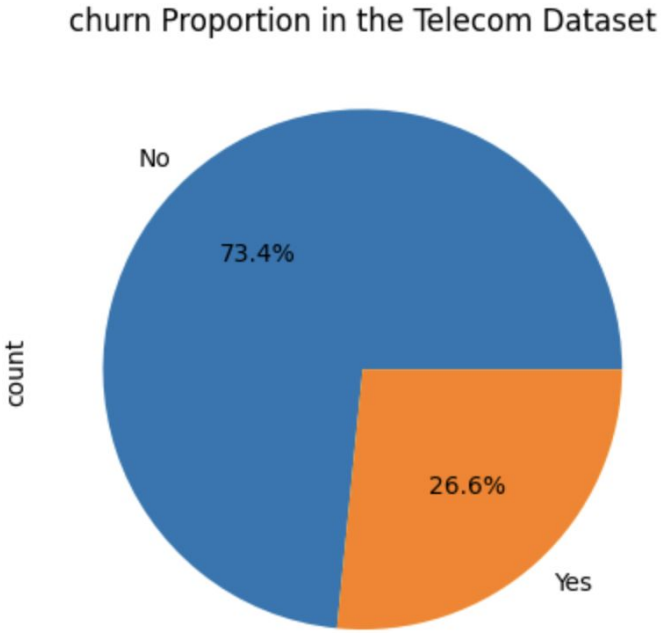
Exploration & Analysis

Sukanya

Churn Proportion

The churn % from the given set of data is 26.6%, which means for every 100 customers, approximately 26.6 members leave the organization.

This in based on the 7043 records



Sukanya

Data Exploration

- The dataset has 21 columns and 7043 clean records
- Target Value is Churn, which has Yes and No Values. It is an object type variable.
- There are no duplicate values
- There are no null values

```
df.shape
```

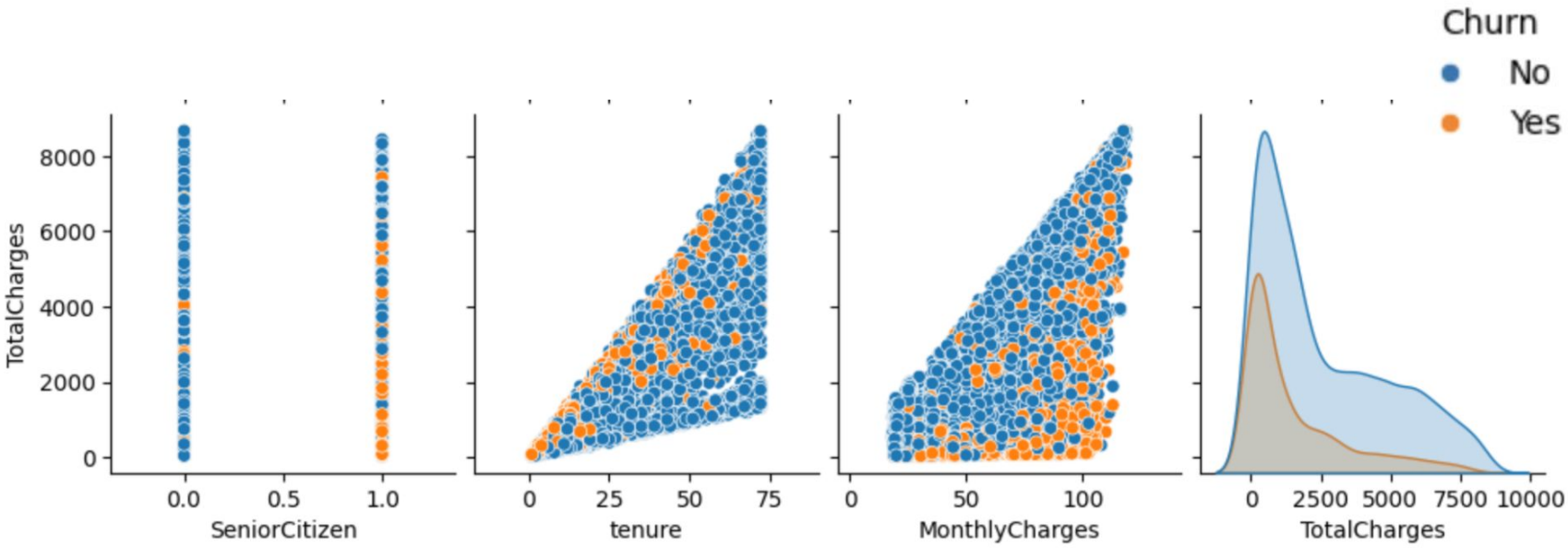
```
(7043, 21)
```

```
#Identifying duplicate records
print(df.duplicated().sum())
```

```
0
```

Sukanya

Relationship between Attributes



Sukanya

Data Preprocessing

```
#identify datatypes
df.dtypes
```

```
customerID           object
gender               object
SeniorCitizen         int64
Partner              object
Dependents           object
tenure                int64
PhoneService         object
MultipleLines        object
InternetService      object
OnlineSecurity       object
OnlineBackup         object
DeviceProtection     object
TechSupport          object
StreamingTV          object
StreamingMovies      object
Contract             object
PaperlessBilling     object
PaymentMethod        object
MonthlyCharges      float64
TotalCharges         object
Churn                object
dtype: object
```

- Total Charges is the target and it is an object datatype
- Changed the object type total charges to numeric
- Any nulls are dropped
- One hot encoding to including categorical variables

Sukanya

Data Wrangling-Data Reduction & Type Conversion

```
# Convert TotalCharges object type to numeric
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
df.dropna(inplace=True)
df.shape
```

```
(7032, 21)
```

```
# Select relevant features (dropping the unused features)
X = df.drop(['customerID', 'Churn'], axis=1)
y = df['Churn']
```

Sukanya

One Hot Encoding

```
# Feature Engineering and Selection
# Encode categorical variables (one-hot encoding)
df = pd.get_dummies(df, columns=['gender', 'Partner', 'Dependents', 'PhoneService',
                                 'MultipleLines', 'InternetService', 'OnlineSecurity',
                                 'OnlineBackup', 'DeviceProtection', 'TechSupport',
                                 'StreamingTV', 'StreamingMovies', 'Contract',
                                 'PaperlessBilling', 'PaymentMethod'])
```

Assigning numerical values to categorical values

Sukanya

Supervised Machine Learning

Sukanya

Model Training

Classification models used:

- KNN
- Logistic Regression
- Random Forest
- Support Vector Machines
- XGBoost

Sukanya

Test-Train

- 80% of data for training
- 20% of data for testing
- StandardScaler used to normalize features
- Applied PCA to the pipeline with 95% variance

```
for name, model in models.items():
    pipe = Pipeline([('scaler', StandardScaler()), ('pca', PCA(n_components=0.95)), (name, model)])
    pipe.fit(X_train_scaled, y_train_encoded)
```

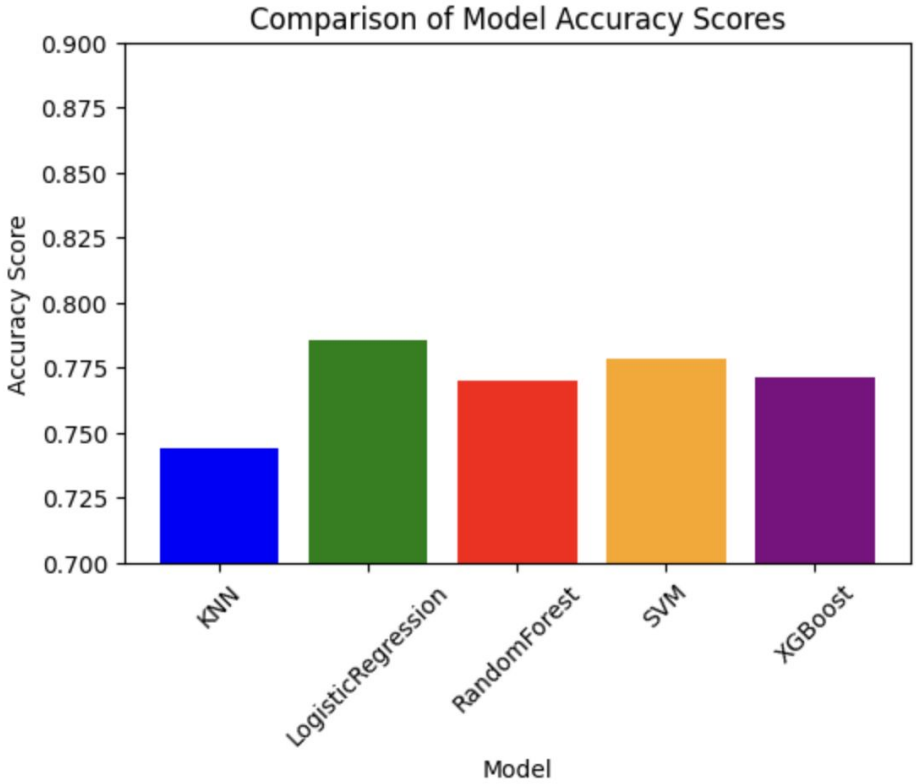
Sukanya

Model Comparison - Classification Report

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1 Score (Churn)
KNN	74.41% 76.55% (HP)	0.57	0.49	0.53
Logistic Regression	78.54%	0.62	0.49	0.49
Random Forest	76.33%	0.59	0.44	0.44
SVM	77.83%	0.61	0.46	0.46
XGBoost	77.11%	0.58	0.51	0.51

Sukanya

Accuracy Comparison



- Logistic Regression has the highest accuracy score and highest precision
- Logistic Regression has the best score in precision

Sukanya

Scope for Improvements

- Increasing data and sample size
- Hyper parameter tuning for all models
  - Grid search to find the best hyper-parameter for other models
- Imbalanced dataset (26% churn from the dataset)
  - Scope for bias

Sukanya

Conclusion

- The calculation of churn greatly helps organizations to take effective customer retention measures.
- Since the customer churn prediction is very sensitive to false negatives, (i.e) the case where the expectation is the customer will retain, while the customer actually leaves, the focus is mainly on accuracy and recall.
- Leveraging insights from the analysis of churn dataset, telecom company can develop and implement customer retention strategies thereby mitigating churn

Sukanya

Thank You