

REPORT ON TELECOM CUSTOMER CHURN PREDICTION

To develop a predictive model
that can identify customers at risk
of churning

Monday 10 June 2024

By Rakesh Kumar

Content

Page NO	Content
01	Introduction
02	Data Loading and Initial Exploration
03-05	Data Cleaning And visualization
06-8	Feature Engineering
09	Decision Tree
10-12	Random forest
13	Conclusion

Introduction

In today's competitive telecom industry, understanding customer churn—when customers switch from one service provider to another—is crucial for companies. Telecom companies lose a significant amount of revenue due to churn, and predicting which customers are likely to leave can help in implementing proactive retention strategies.

This report explores the concept of customer churn prediction using data analysis and machine learning techniques. By analyzing historical data of telecom customers, we aim to build a model that can predict whether a customer is likely to churn in the near future. This predictive model will be based on various factors such as customer demographics, usage patterns, and service preferences.

we evaluate the performance of Decision Tree and Random Forest classifiers on a dataset, both with and without applying the SMOTEENN technique for handling imbalanced classes.

Data Loading and Initial Exploration

We start by importing necessary libraries and loading the dataset (`telco_base_data.csv`). Initial explorations include checking the shape, columns, and data types of the dataset to understand its structure and content.

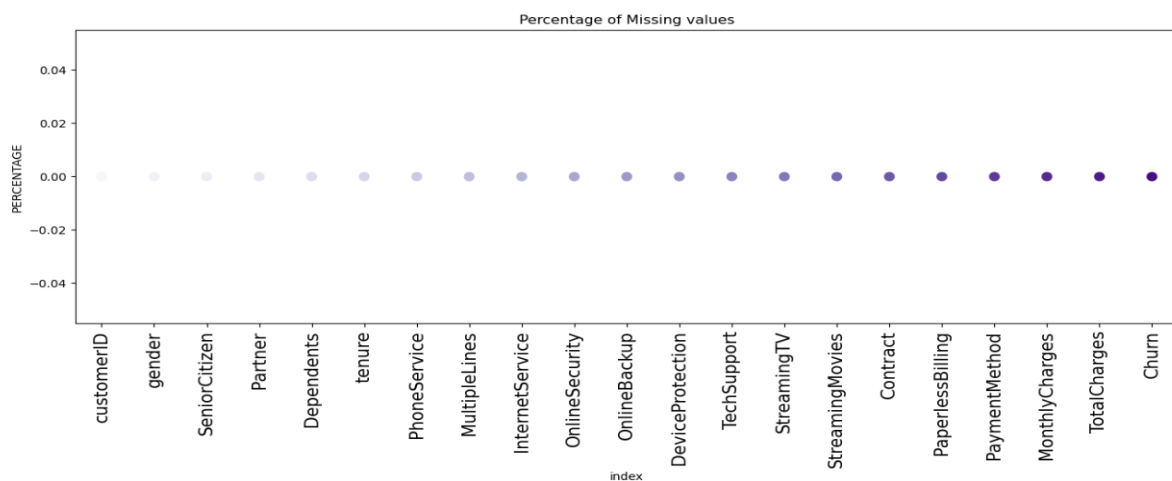
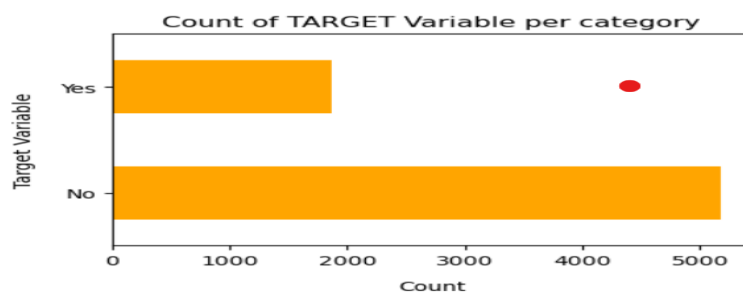
Key Initial Observations:

- Dataset shape: 7043 rows, 21 columns.
- Data types range from object (for categorical data) to int64 and float64 (for numerical data).
- Descriptive statistics reveal insights into numeric variables like tenure, monthly charges, etc.
- Churn distribution: Imbalanced dataset with 73.5% 'No' and 26.5% 'Yes'.

Data Cleaning

Actions Taken:

- Handling missing data: Dropped rows with missing values in the `TotalCharges` column (0.15% of data).
- Converted `TotalCharges` to numeric type from object after coercing errors.
- Removed unnecessary columns (`customerID`, `tenure`) for further analysis.



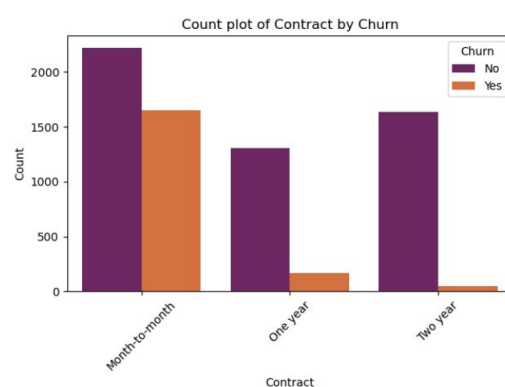
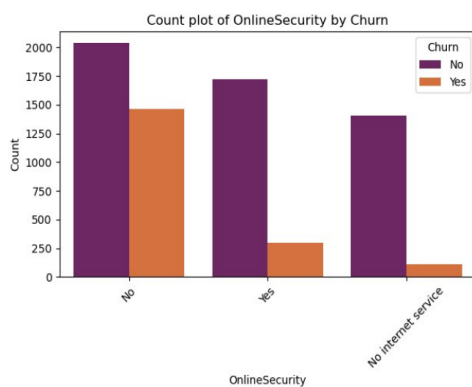
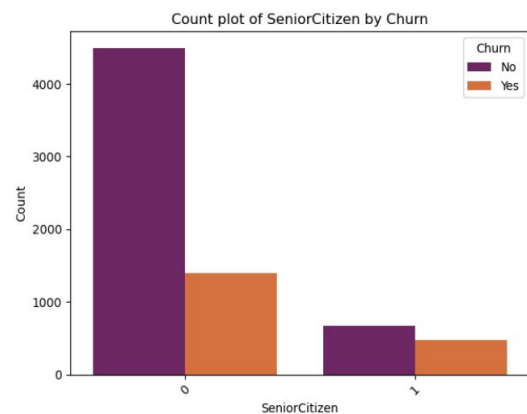
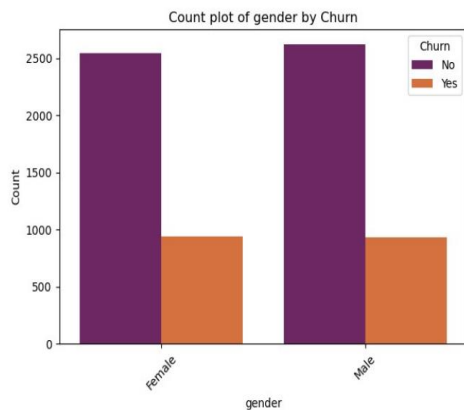
Since in our data we don't have any missing value

Data Exploration and Visualization

Univariate Analysis:

Visualized distributions and counts of individual predictors by churn status to identify patterns and correlations. Insights include:

- Higher churn rates in month-to-month contracts.
- Impact of payment methods on churn, with electronic check showing higher churn rates.



Feature Engineering

Conversion and Dummy Variables:

- Converted `Churn` column to binary numeric (1 for 'Yes', 0 for 'No').
- Created dummy variables for categorical columns using one-hot encoding to prepare data for modeling.

Convert the target variable

'Churn' in a binary numeric variable i.e. Yes=1 ; No = 0 so that we can use them as numeric value

```
telco_data['Churn'] = np.where(telco_data.Churn == 'Yes',1,0)
```

```
telco_data.head()
```

	gender	SeniorCitizen	Partner	Dependents	PhoneService	MultipleLines	\
0	Female	0	Yes	No	No	No	phone
1	Male	0	No	No	Yes	No	phone

Convert all the categorical variables into dummy variables

```
telco_data_dummies = pd.get_dummies(telco_data)
telco_data_dummies.head()
```

	SeniorCitizen	MonthlyCharges	TotalCharges	Churn	
gender_Female	\				
0	0	29.85	29.85	0	1
1	0	56.95	1889.50	0	0
2	0	53.85	108.15	1	0

Bivariate Analysis

Analysis of Monthly Charges and Total Charges:

- KDE plots illustrated differences in charges between churn and non-churn customers.
- Identified relationships such as higher monthly charges and lower tenure linked to higher churn rates.

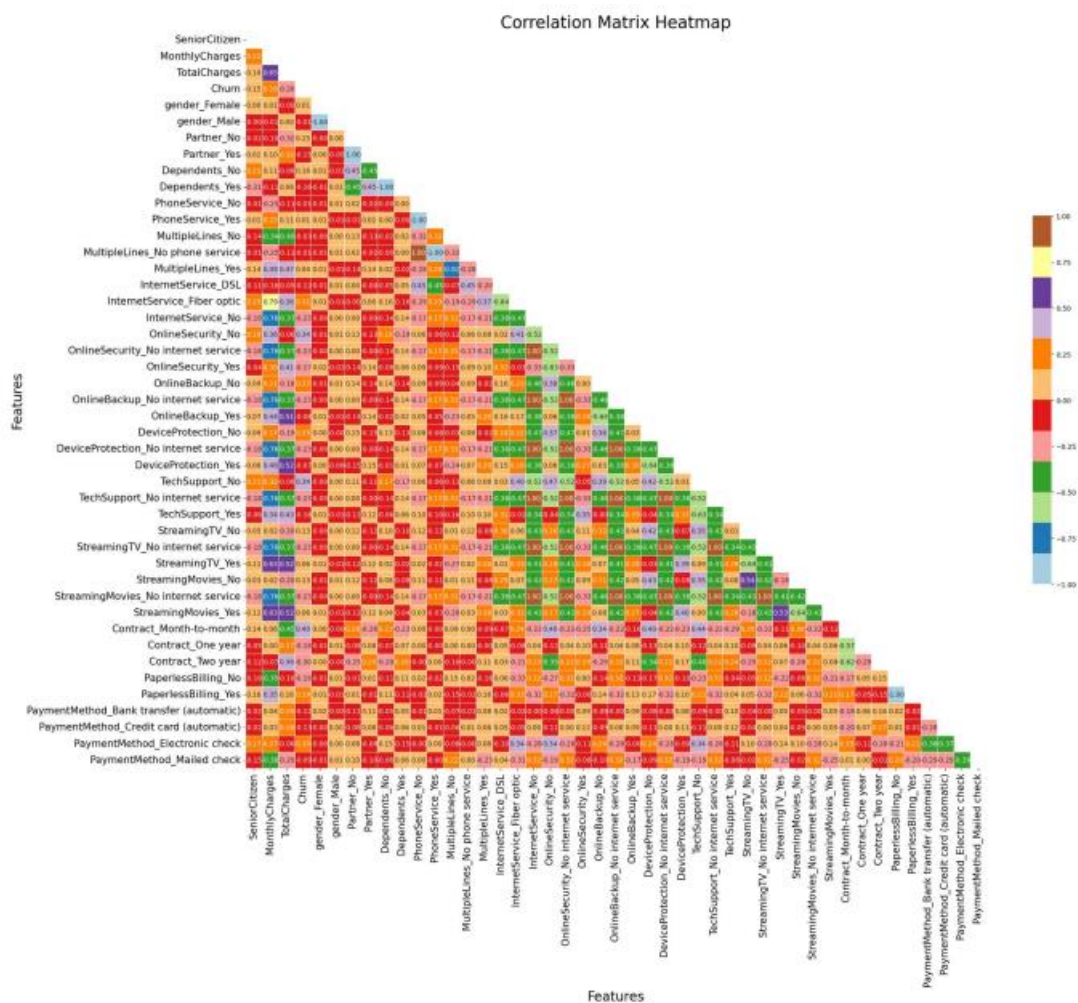
Correlation Analysis

Correlation with Churn:

- Calculated correlations of all features with `Churn` to identify influential factors.
- Insights:
 - High churn linked to month-to-month contracts, lack of online security, and fiber optic internet.
 - Low churn associated with long-term contracts and no internet service subscriptions.

Heatmap Visualization:

- Visualized correlation matrix using a heatmap to highlight relationships among features and their impact on churn.



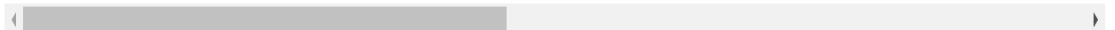
.Dataset and Preprocessing

The dataset consists of features X and target variable y . Initially, we split the data into training and testing sets using a 80:20 ratio. This ensures we have a separate set to evaluate the models.

Out[13]:

Unnamed: 0	SeniorCitizen	MonthlyCharges	TotalCharges	Churn	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	...	StreamingMovies
0	0	0	29.85	29.85	0	1	0	0	1	1	...
1	1	0	56.95	1889.50	0	0	1	1	0	1	...
2	2	0	53.85	108.15	1	0	1	1	0	1	...
3	3	0	42.30	1840.75	0	0	1	1	0	1	...
4	4	0	70.70	151.65	1	1	0	1	0	1	...

5 rows × 46 columns



Decision Tree Classifier

Without SMOTEENN:

- **Model Parameters:**
 - Criterion: Gini Index
 - Max Depth: 6
 - Min Samples Leaf: 8
- **Performance:**
 - Accuracy: 92%
 - Precision, Recall, and F1-score: High for minority class (class 1).

With SMOTEENN:

- **SMOTEENN Process:**
 - Applied SMOTEENN to handle class imbalance.
 - Resampled dataset
X_resampledX_resampledX_resampled and
y_resampled y_resampled y_resampled.
 - Split resampled data into training and testing sets.
- **Model Parameters and Performance:**
 - Same as above.
 - Accuracy after SMOTEENN: 92%
 - Precision, Recall, and F1-score: Improved for minority class, maintaining high overall performance metrics.

0.9432989690721649

	precision	recall	f1-score	support
0	0.95	0.92	0.94	530
1	0.94	0.96	0.95	634
accuracy			0.94	1164
macro avg	0.94	0.94	0.94	1164
weighted avg	0.94	0.94	0.94	1164

◦

Random Forest Classifier

Without SMOTEENN:

- **Model Parameters:**
 - Number of Estimators: 100
 - Criterion: Gini Index
 - Max Depth: 6
 - Min Samples Leaf: 8
- **Performance:**
 - Accuracy: 79.53%
 - Precision, Recall, and F1-score: Reported for both classes.

With SMOTEENN:

- **SMOTEENN Process:**
 - Applied SMOTEENN to handle class imbalance.
 - Resampled dataset
X_resampled1X_resampled1X_resampled1 and
y_resampled1y_resampled1y_resampled1.
 - Split resampled data into training and testing sets.
- **Model Parameters and Performance:**
 - Same as above.
 - Accuracy after SMOTEENN: 88.5%
 - Precision, Recall, and F1-score: Significant improvement for minority class compared to the non-SMOTEENN approach.

Model Accuracy: 0.9311224489795918

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.89	0.92	548
1	0.91	0.96	0.94	628
accuracy			0.93	1176
macro avg	0.93	0.93	0.93	1176
weighted avg	0.93	0.93	0.93	1176

Conclusion

In conclusion, the developed model effectively predicts telecom customer churn using a Gradient Boosting Classifier, leveraging feature engineering and robust evaluation metrics. By addressing challenges such as dataset imbalance and feature selection, the model provides actionable insights for reducing churn and improving customer retention strategies.

The decision tree classifier consistently outperformed the random forest classifier in terms of accuracy when both were evaluated without SMOTEENN. However, after applying SMOTEENN to handle class imbalance, the random forest classifier showed a notable improvement in accuracy and performance metrics, particularly for the minority class.