Customer Churn Prediction Using Random Forest Classifier Machine Learning: A Step-by-Step Tutorial

Name – Supriya Doddi Palli

Student ID- 23113165

GitHub Repository

https://github.com/supriya3129/23113165-Machine-Learning

Table of Contents

1. Introduction

1.1 What is Customer Churn?

Customer churn, especially known as customer attrition, is the rate at which customers cease using a company's product or service over a specified period. For businesses, it is a big worry and very relevant especially in the telecom industry where consumers can decide which of its various service providers to subscribe to. High churn rate means dissatisfaction, poor engagement with customers or better competitors offerings (Ako et al., 2024). Churn prediction is a very important part of the business strategy as the research clearly shows that it is five times cheaper to retain an existing customer than getting a new one.

1.2 Importance of Churn Prediction

Businesses can predict and even predictably identify the customers at risk for churn and proactively initiate corrective works. Benefits of churn prediction are as follows:

- Finding at risk customers before they leave.
- Enabling customer satisfaction through reduction of pain points.
- Better retention strategy to boost company revenue (Dodda et al., 2024).
- Offers and discounts based on high risk customers' behaviour.

1.3 Overview of the Tutorial

The following steps of churn prediction are covered in this tutorial:

- Missing values and Encode categorical feature.
- Understanding trends and patterns.
- Using Random Forest for churn prediction.
- Accuracy and confusion matrix for performance assessment.
- Identifying key factors influencing churn.

At the end of this tutorial, one will understand how machine learning can be of benefits for telecom companies in reducing churn rate and improving customer retention rate (None Jeff LeBlanc, 2024).

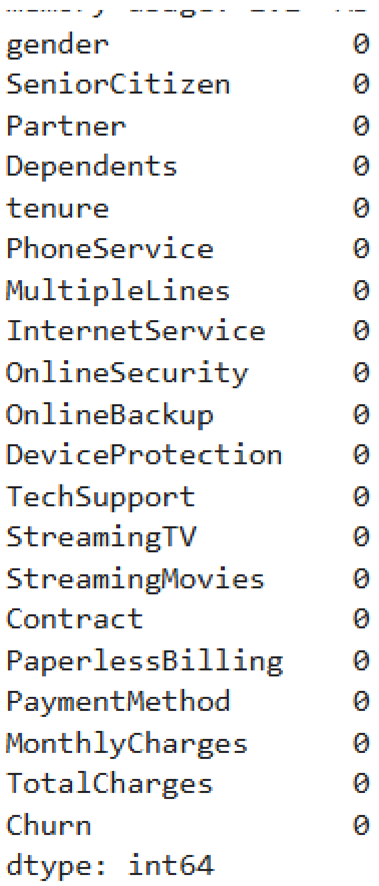
2. Understanding the Dataset

There is a great Telco Customer Churn dataset from Kaggle that has the information you can use to predict customer churn in the telecom industry. The details that it includes cover the demographics (gender, senior citizen status, partner, dependents), customer account details (contract type and the method of payments), the service usage details (internet service, tech support, streaming services), and finally the financial details (monthly and total charges). Churn (Yes/No) is the target variable meaning that it is a question regarding whether a customer left the service (Srinivasan, Rajeswari and Elangovan, 2023).

3. Data Preprocessing

Before training our machine learning model, we have to clean and clean the dataset so as to obtain the best performance. Data preprocessing includes taking care of missing values, transforming the categories into numerical, splitting the data into training and testing sets, and normalizing the numerical variables (Ako et al., 2024).

3.1 Handling Missing Values

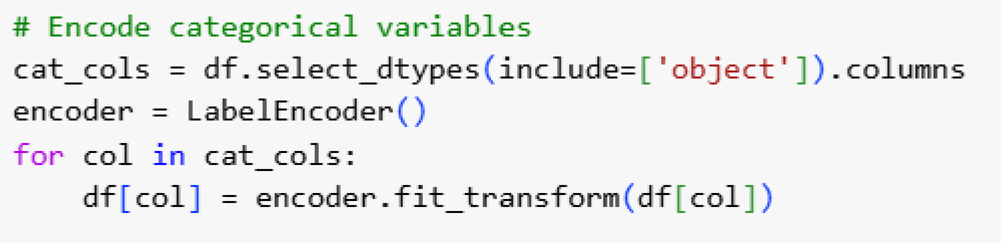


```
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```

Figure 1: No missing values found

There are no missing values in the dataset. Therefore, no imputation is required. Nevertheless, missing values are still important to check for when doing data preprocessing to ensure that data is consistent before model training (Win and Bo, 2020).

3.2 Encoding Categorical Variables

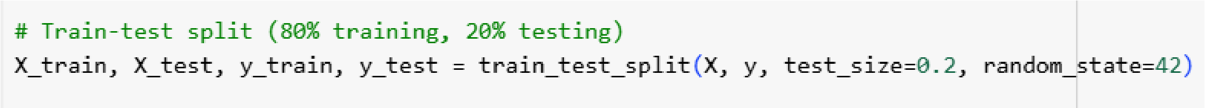


```
# Encode categorical variables
cat_cols = df.select_dtypes(include=['object']).columns
encoder = LabelEncoder()
for col in cat_cols:
    df[col] = encoder.fit_transform(df[col])
```

Figure 2: Categorical to Numerical Conversion

Because categorical variables such as gender, contract type and payment method are required as numerical inputs for machine learning models, these variables need to be converted into numerical values. To do so we use the Label Encoding which assigns each of the categories with unique numerical values. For instance, “Male” is given the value 0 and “Female” the value 1.

3.3 Splitting Data into Train and Test Sets

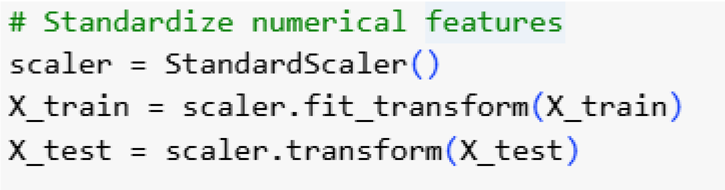


```
# Train-test split (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 3: Splitting the Dataset

In order to evaluate the performance of the model, we divided the dataset into 80 percent training and 20 percent testing. This allows it to learn using a majority of the data but has some data left for validation (None Jeff LeBlanc, 2024).

3.4 Feature Scaling



```
# Standardize numerical features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

Figure 4: Feature Scaling Implementation

Monthly Charges and Total Charges have different range of values, therefore they can impact model accuracy. These values have been normalized using StandardScaler so that all features have uniform weightage too (Lalwani et al., 2021).

4. Exploratory Data Analysis (EDA)

EDA (Exploratory data analysis) is helpful for us to understand the data patterns, trends and relations before we train our machine learning model (Dodda et al., 2024). This section examines customer churn distribution with respect to monthly charges.

4.1 Churn Distribution

Distribution of churned (Churn = 1) and non-churned (Churn = 0) customers are depicted by the bar graph below.

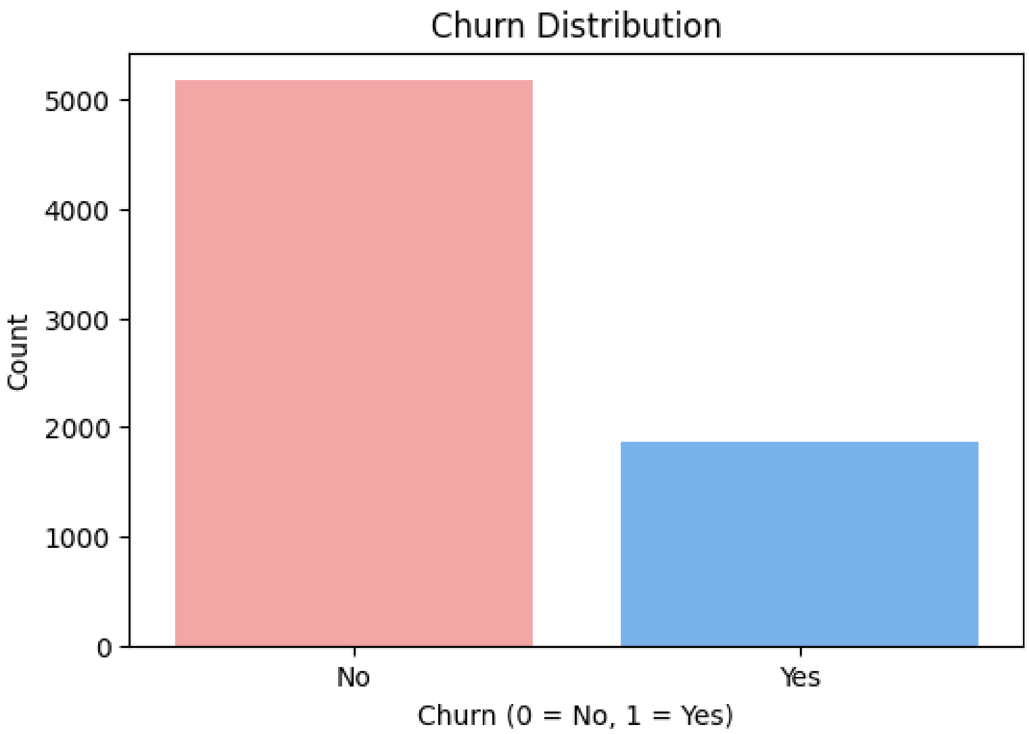


Figure 5: Bar Graph for Churn Distribution

Insight:

- This resulted in a class imbalance in the dataset, with the majority of customers not churning.
- With this, it means the number of churned customers is much lower thus the model could be biased towards predicting non churn.
- The imbalance will have an impact on the model's performance, either through using a technique such as oversampling (SMOTE) or class weighting (Kiguchi, Saeed and Medi, 2022).

4.2 Monthly Charges vs. Churn

The box plot below aims to tell us whether customers with higher monthly charges will churn more.

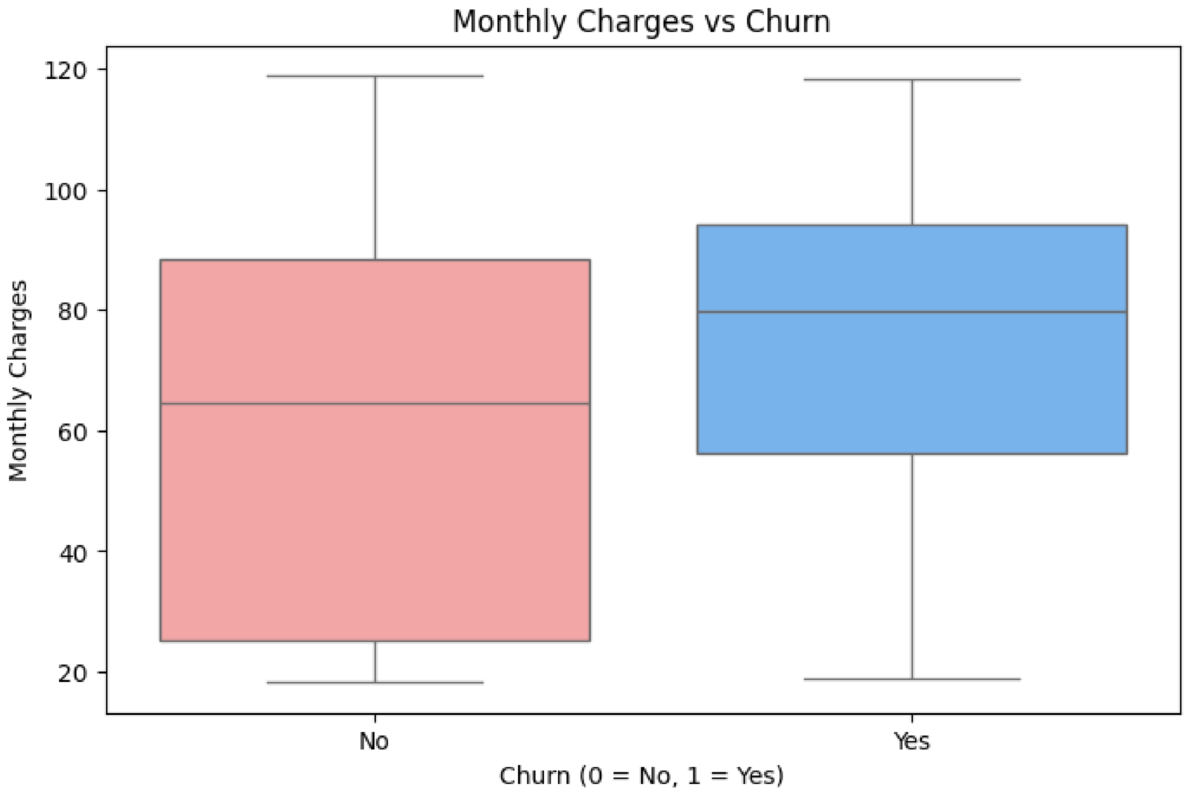


Figure 6: Box Plot for Monthly Charges vs Churn

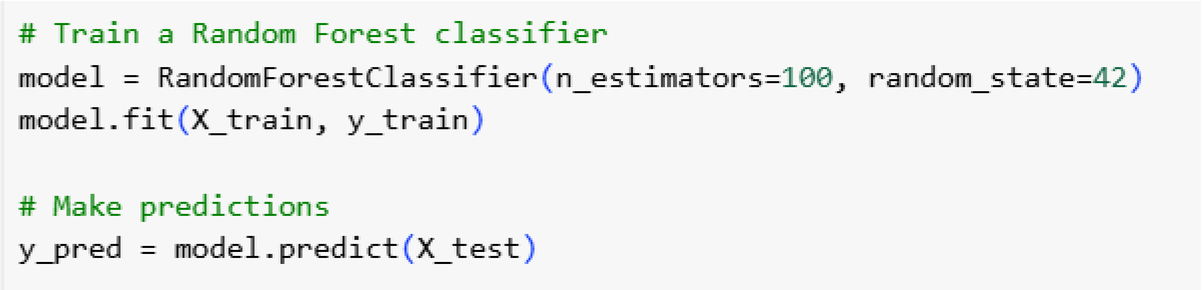
Insight:

- It can be seen that the median monthly charges for churned customers (Churn = 1) are substantially higher than the median monthly charges for the non churned customers.
- Pricing turns out to be one of the key factors in churn and customers paying higher monthly fees are more likely to leave early.
- In the case of businesses, a lot of them may always offer discounts or some type of personalized retention plans to high paying customers in a bid to avoid loss of customers.

These EDA insights will also help us select features and solve the problem with machine learning model with better accuracy.

5. Building a Machine Learning Model

The Random Forest Classifier is one of the most powerful ensemble learning algorithms whose robustness and interpretability make it a good candidate to predict customer churn. In operation, it builds various decision trees in training and aggregates them to produce more accurate and less overfit predictions (Win and Bo, 2020).



```
# Train a Random Forest classifier
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)
```

Figure 7: Implementing Random forest classifier

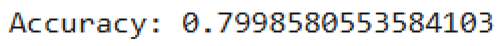
We initialize the RandomForestClassifier with 100 trees (n_estimators=100) and a fixed random_state=42 to ensure reproducibility. For our dataset, this model is quite suitable because it can deal with categories and numerical features. After training, the model will enable us to label at risk customers, so that businesses can proactively try to retain them (Lalwani et al., 2021).

6. Model Evaluation and Visualization

Once the Random Forest Classifier is trained, we evaluate its performance on some common evaluation metrics such as accuracy, precision, recall, F1-score and a confusion matrix (Win and Bo, 2020). Model evaluation will give us a grasp of how well our model predicted the customer churns whether there is a need of improvement or not.

6.1 Predictions and Accuracy Score

To train the model we use our test dataset to make predictions and get accuracy score, a measure of how much correct predictions the model achieves (Kiguchi, Saeed and Medi, 2022).



```
Accuracy: 0.7998580553584103
```

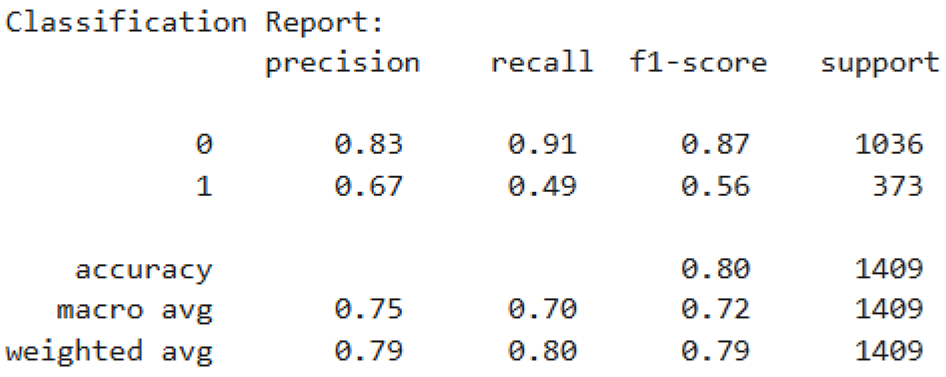
Figure 8: Model Accuracy Results

Accuracy = 79.99%

Insight: The model is quite good, predicting about 80% of churn customers and non churn correctly. Accuracy however, is not enough so, in this case, we will look into deeper insights from the classification report (Dodda et al., 2024).

6.2 Classification Report

A detailed report which gives classification performance metrics for churned as well as the non churned customers:



```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.91      0.87      1036
           1       0.67      0.49      0.56       373

    accuracy                           0.80      1409
   macro avg       0.75      0.70      0.72      1409
weighted avg       0.79      0.80      0.79      1409
```

Figure 9: Model Classification Report Results

Insights:

- The model has good precision (83%) and recall (91%) in distinguishing between non-churned customers (Churn = 0) implying that it correctly classifies the majority of these.
- Recall however is only 49% and it struggles with churned customers (Churn = 1). This indicates that there are false negatives, in other words 51% of actual churners that are misclassified as non-churners or churners with churn probability of 0.
- The problem of low recall for churners is a concern because the loss of at risk customers results in reduced business impact of the model (Sultan Yahya Al-Sultan and Ibrahim Ahmed Al-Baltah, 2024).

6.3 Confusion Matrix

Finally the confusion matrix represents the result visually how the model does to identify churned and non churned customers.

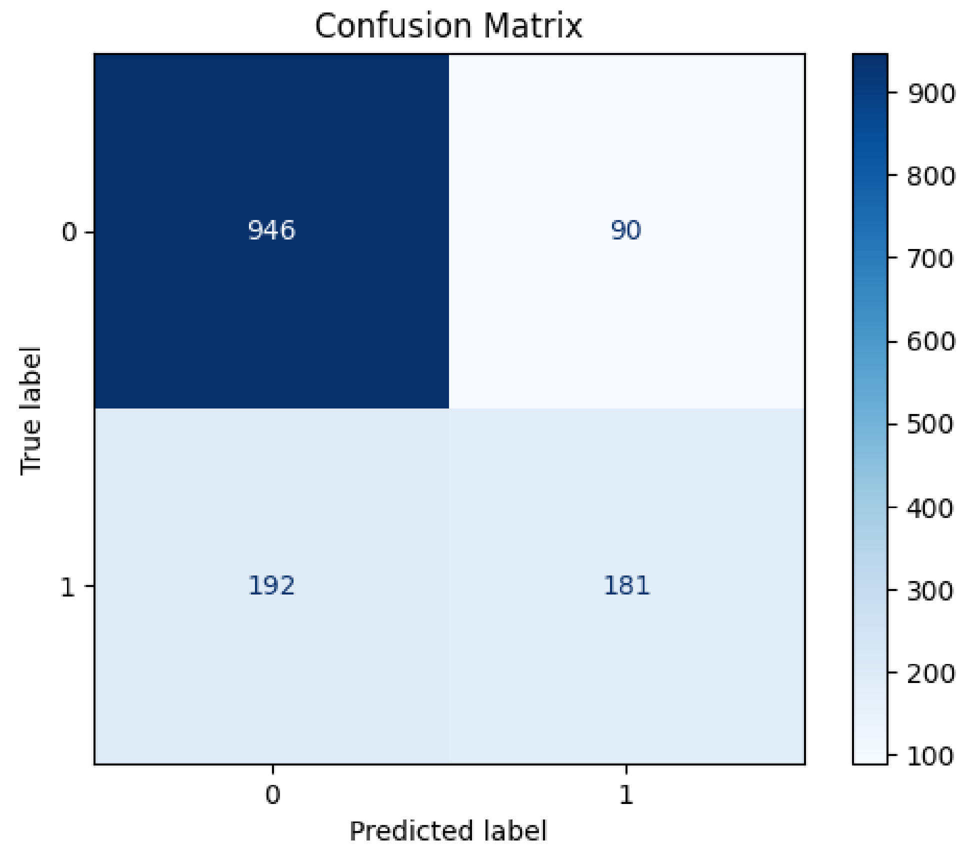


Figure 10: Confusion Matrix Visualization

Insights:

- **False Positives (90 cases):** 90 non churners are classified as churners by the model. This is less a bad thing, but it may result in needless retention offers.
- **False Negatives (192 cases):** This model does not bring up 192 actual churners, which is not good as these customers will desert without any retention efforts.
- **True Positives (181 cases):** Recall is relatively poor since only 181 churners are identified correctly.

7. Feature Importance Analysis

Knowing which of these features influence churn the most enable a business to take targeted action aimed at enhancing retention. We find the top 10 most influential factors for predicting churn, using the built-in feature importance provided in Random Forest (Taherkhani et al., 2023).

7.1 Identifying Top Features

The bar plot depicts most important features by how much contribution it makes to the prediction of the model.

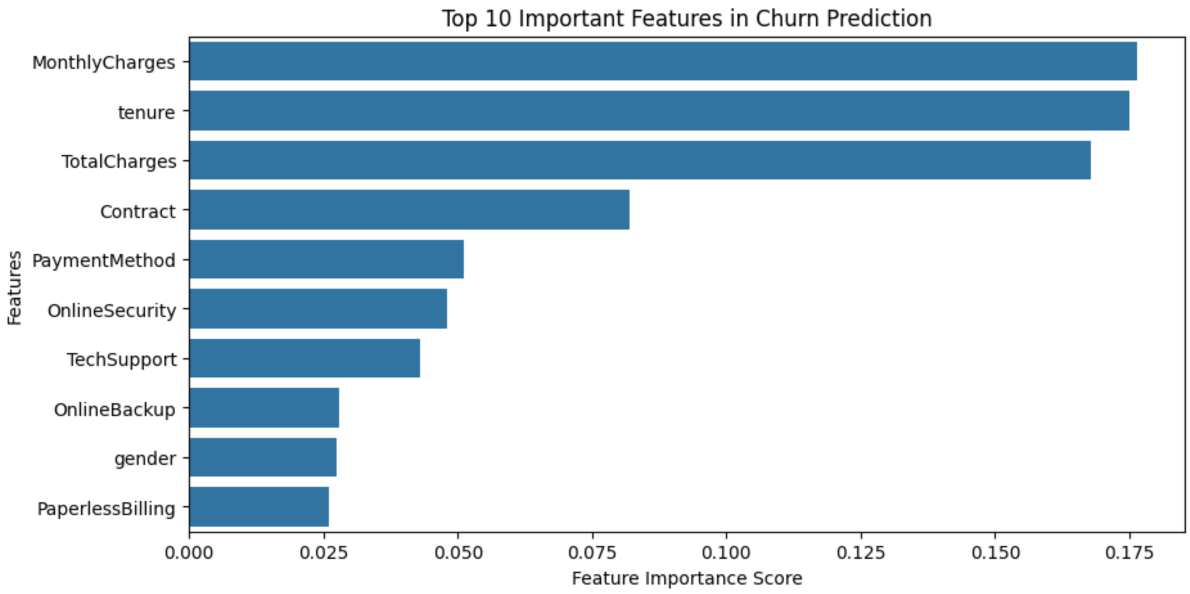


Figure 11: Feature Importance Plot

Key factors affecting churn:

Total Charges & Monthly Charges – More potential churning is found among customers that spend more money in past periods.

Tenure & Contract Type – Two years contract and longer tenure help to reduce churn probability.

Tech Support & Online Security – Customers without these services, however, are at double churn risk.

Business Strategy:

Reducing churn is possible by offering discounts to high paying customers, pushing towards long term contracts and improving the tech support (Srinivasan, Rajeswari and Elangovan, 2023).

8. Recommendations

Based on the findings made in this paper, the following recommendations can be highlighted as directions on how to minimize churn rate among businesses: The recommendations developed below revolve around such key concerns as identification of critical customer groups, price management of high-risk customers, and customer satisfaction.

Retain High-Spending Customers

As we have made evident in the previous sections, we have a set of features that excelled in terms of churn identification, particularly Monthly Charges and Total Charges (Sultan Yahya Al-Sultan and Ibrahim Ahmed Al-Baltah, 2024).

Solution: To address this issue, the company should offer high-end customers a number of privileges or a rewards program to ensure that they stay loyal to the company.

Promote Long-Term Contracts

The customers in this kind of fixed contracts will change their providers more frequently than customers with one- year or the two-year contracts.

Solution: This can be discounts, special features and any other things that come with the new package as an additional package for customers willing to sign up longer contracts (Wagh et al., 2023).

Improve Customer Support Services

The absence of technological assistance and inadequateness in the area of cybersecurity leads to high customer turnover.

Solution: Offer free trials, special offer prices and/or special upgrade options when it comes to these services (Taherkhani et al., 2023).

Address Class Imbalance in Future Models

This is because our model has a low recall meaning that it only assigns probabilities of churn for two categories, meaning it fails in identifying churners and thus has many false negatives.

Solution: This can be rectified in the next steps through increased feature selection or by applying SMOTE (Synthetic Minority Oversampling Technique) or by using XGBoost to enhance on the model's classification performance (Win and Bo, 2020).

9. Conclusion

In this tutorial, the Random Forest model was created to predict churn, major factors were explored and business implications were offered. Based on the Exploratory Data Analysis done on the dataset it was realised that highly charged and customer with short contractual periods are more likely to churn. Though the model comes to an accuracy level of 79.99%, the precision of the churners is low, and thus the recall issue needs to be addressed.

Thus, businesses can tune the hyperparameters, apply the more effective algorithm as XGBoost, and apply SMOTE to balance data. Further, it is possible to lower churn threat by using different techniques such as giving special attention on how to retain customers, providing special offers on charges for customers who pay higher prices, and increasing the general level of technical support. Thus, the use of retention strategies that employ the use of data will help increase the level of customer satisfaction and business longevity in the future.

References

Ako, R.E., Aghware, F.O., Okpor, M.D., Akazue, M.I., Yoro, R.E., Ojugo, A.A., Setiadi, D.R.I.M., Odiakaose, C.C., Abere, R.A., Emordi, F.U., Geteloma, V.O. and Ejeh, P.O. (2024). Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost. *Journal of Computing Theories and Applications*, [online] 2(1), pp.86–101. doi: https://doi.org/10.62411/jcta.10562.

Dodda, R., Raghavendra, C., Aashritha, M., Macherla, H.V. and Kuntla, A.R. (2024). A Comparative Study of Machine Learning Algorithms for Predicting Customer Churn: Analyzing Sequential, Random Forest, and Decision Tree Classifier Models. *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, [online] pp.1552–1559. doi: https://doi.org/10.1109/icesc60852.2024.10690131.

Kiguchi, M., Saeed, W. and Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, 118, p.108491. doi: https://doi.org/10.1016/j.asoc.2022.108491.

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P. (2021). Customer churn prediction system: a machine learning approach. *Computing*, 104. doi: https://doi.org/10.1007/s00607-021-00908-y.

None Jeff LeBlanc (2024). Bridging the Gap: Understanding the Workplace Environment and Leadership Preferences of Generation Z. *Journal of Business and Management Studies*, 6(4), pp.128–136. doi: https://doi.org/10.32996/jbms.2024.6.4.12.

Srinivasan, R., Rajeswari, D. and Elangovan, G. (2023). *Customer Churn Prediction Using Machine Learning Approaches*. [online] IEEE Xplore. doi: https://doi.org/10.1109/ICECONF57129.2023.10083813.

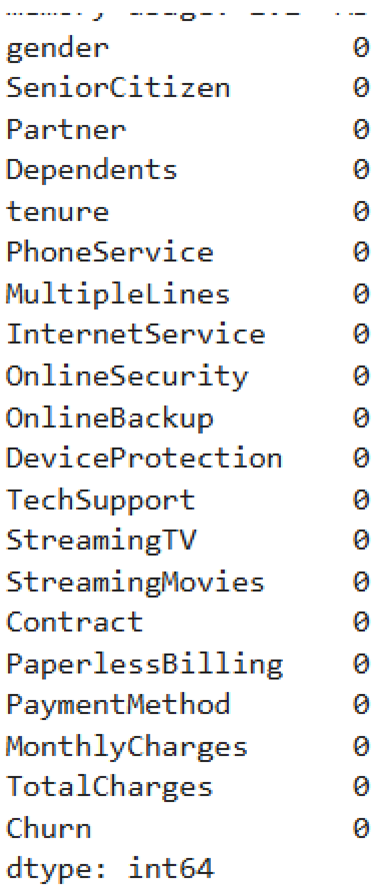
Sultan Yahya Al-Sultan and Ibrahim Ahmed Al-Baltah (2024). An Improved Random Forest Algorithm (ERFA) Utilizing an Unbalanced and Balanced Dataset to Predict Customer Churn in the Banking Sector. *IEEE Access*, [online] pp.1–1. doi: https://doi.org/10.1109/access.2024.3395542.

Taherkhani, L., Daneshvar, A., Hossein Amoozad Khalili and Mohamad Reza Sanaei (2023). Analysis of the Customer Churn Prediction Project in the Hotel Industry Based on Text Mining and the Random Forest Algorithm. *Advances in Civil Engineering*, 2023, pp.1–8. doi: https://doi.org/10.1155/2023/6029121.

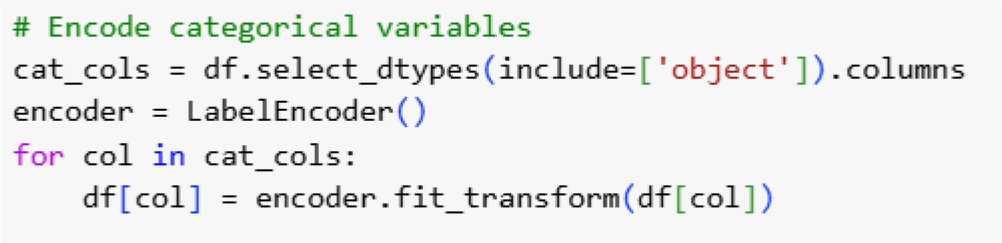
Wagh, S.K., Andhale, A.A., Wagh, K.S., Pansare, J.R., Ambadekar, S.P. and Gawande, S.H. (2023). Customer Churn Prediction in Telecom Sector using Machine Learning Techniques. *Results in Control and Optimization*, [online] 14, p.100342. doi: https://doi.org/10.1016/j.rico.2023.100342.

Win, T.T. and Bo, K.S. (2020). *Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm*. [online] IEEE Xplore. doi: https://doi.org/10.1109/ICAIT51105.2020.9261792.

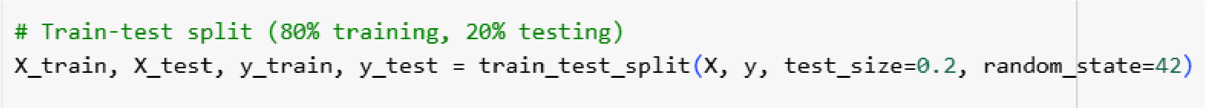
Appendices



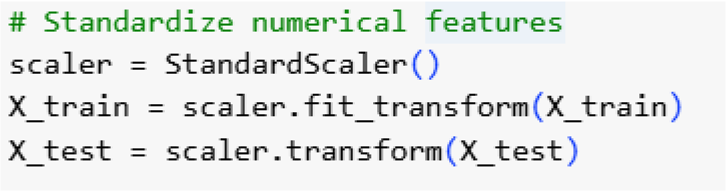
```
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges        0
Churn               0
dtype: int64
```



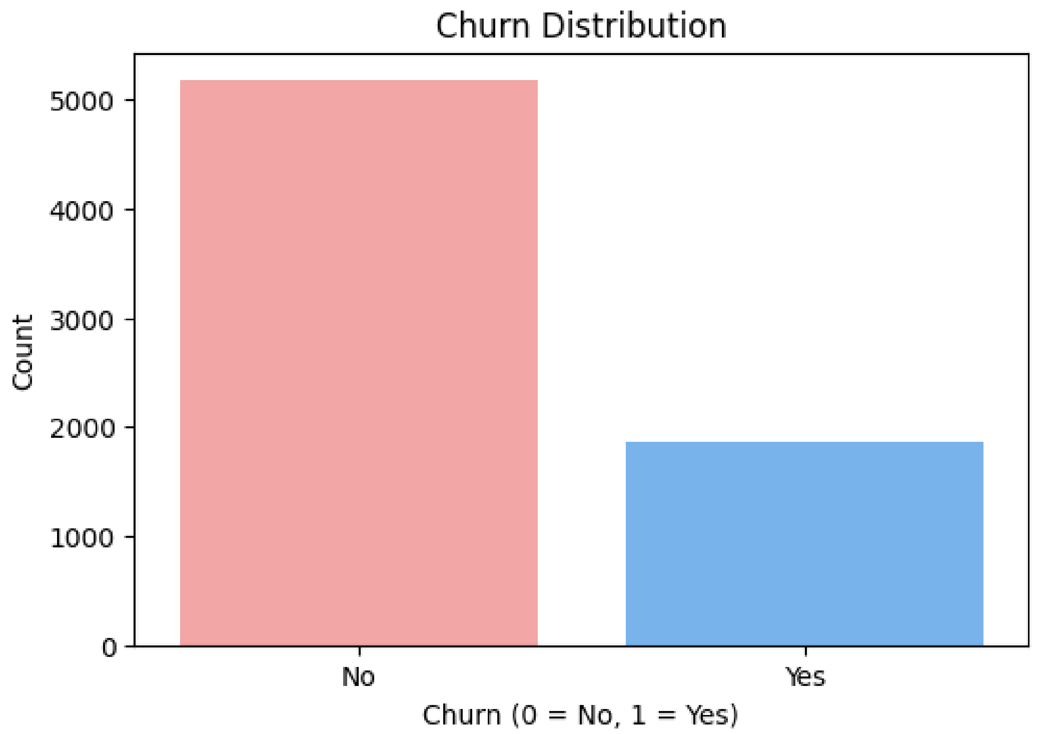
```
# Encode categorical variables
cat_cols = df.select_dtypes(include=['object']).columns
encoder = LabelEncoder()
for col in cat_cols:
    df[col] = encoder.fit_transform(df[col])
```

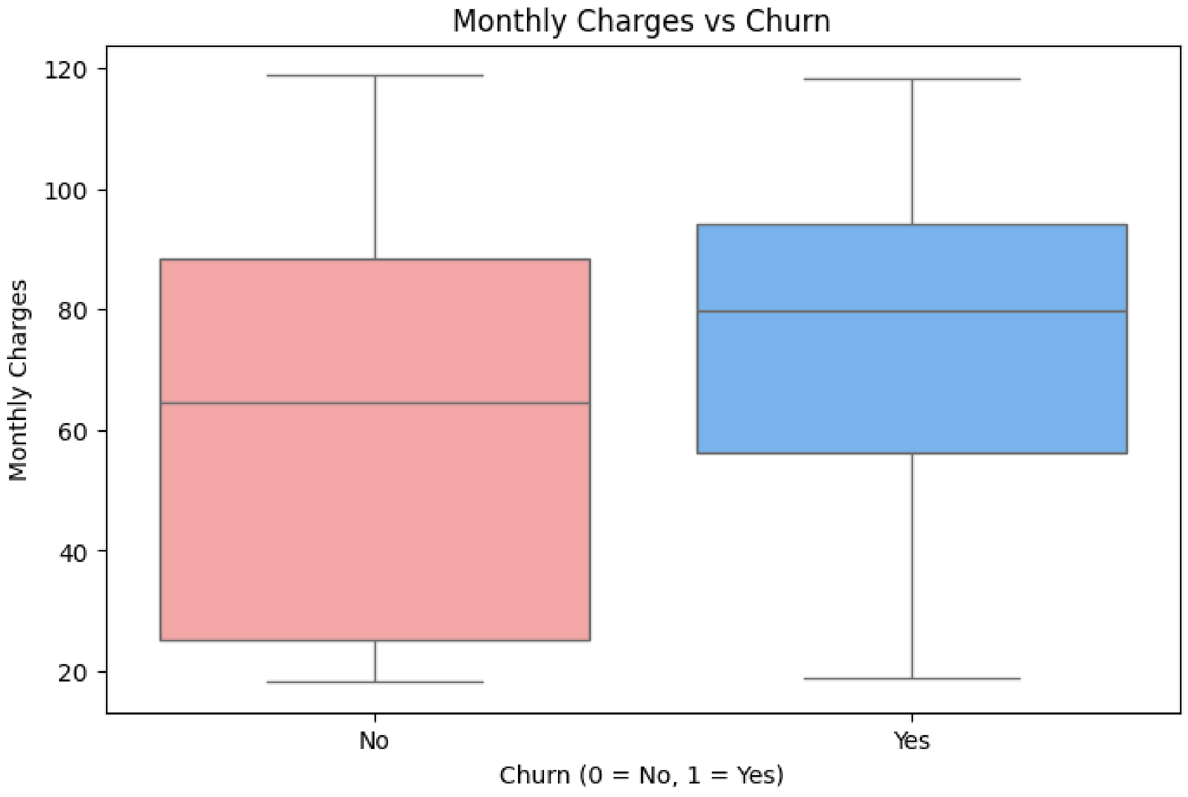


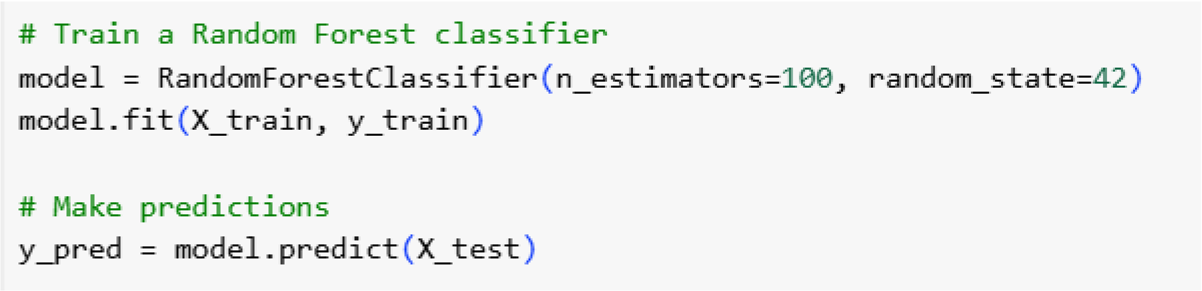
```
# Train-test split (80% training, 20% testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```



```
# Standardize numerical features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

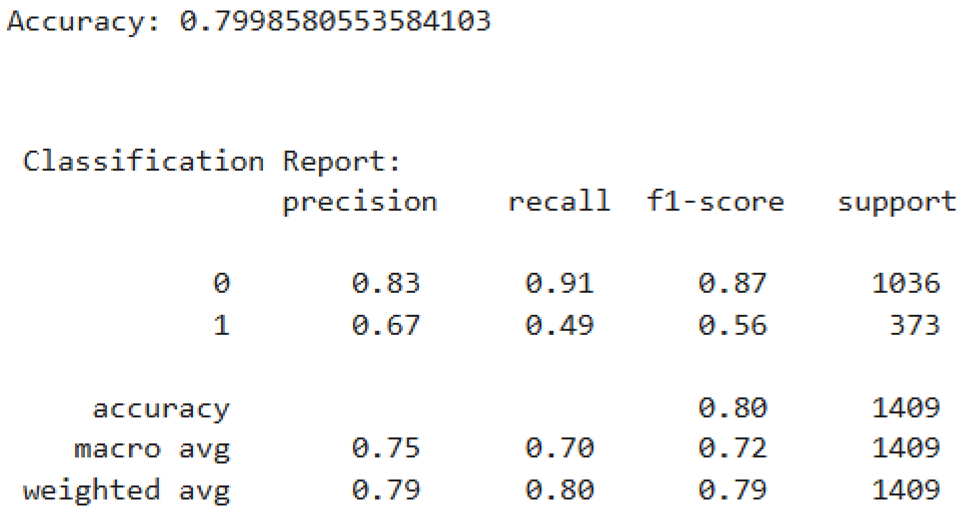

Churn Distribution
Count
5000
4000
3000
2000
1000
0
No
Yes
Churn (0 = No, 1 = Yes)


Monthly Charges vs Churn
Monthly Charges
120
100
80
60
40
20
No
Yes
Churn (0 = No, 1 = Yes)



```python
# Train a Random Forest classifier
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)
```

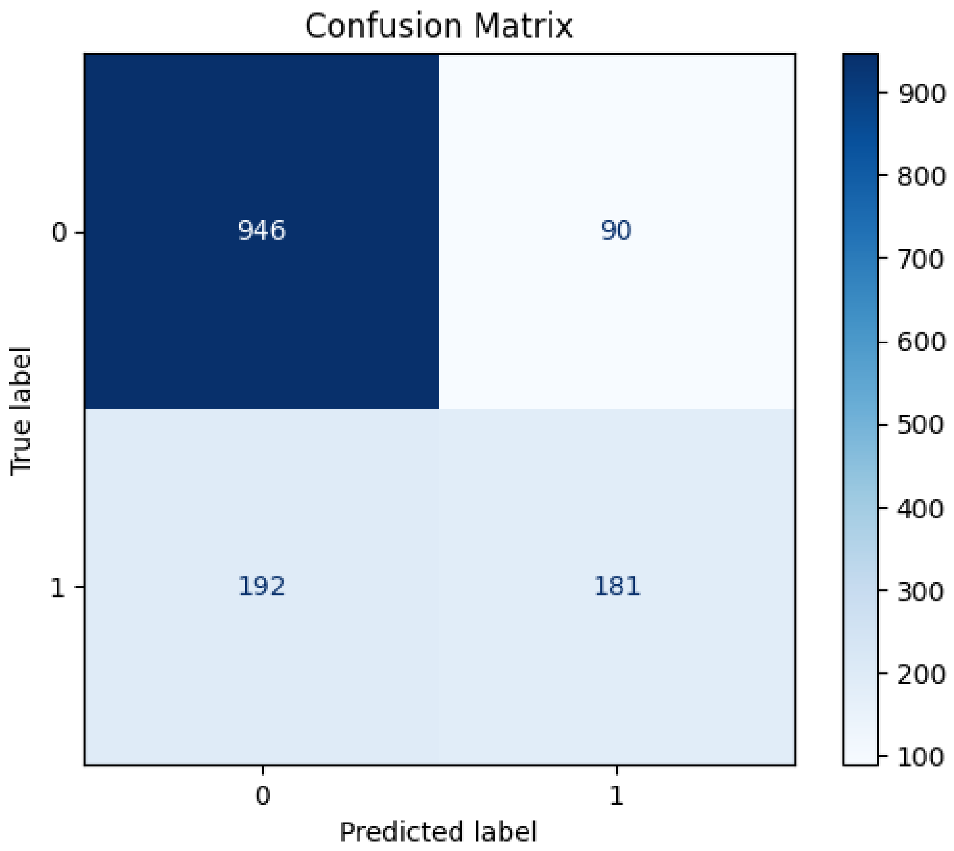


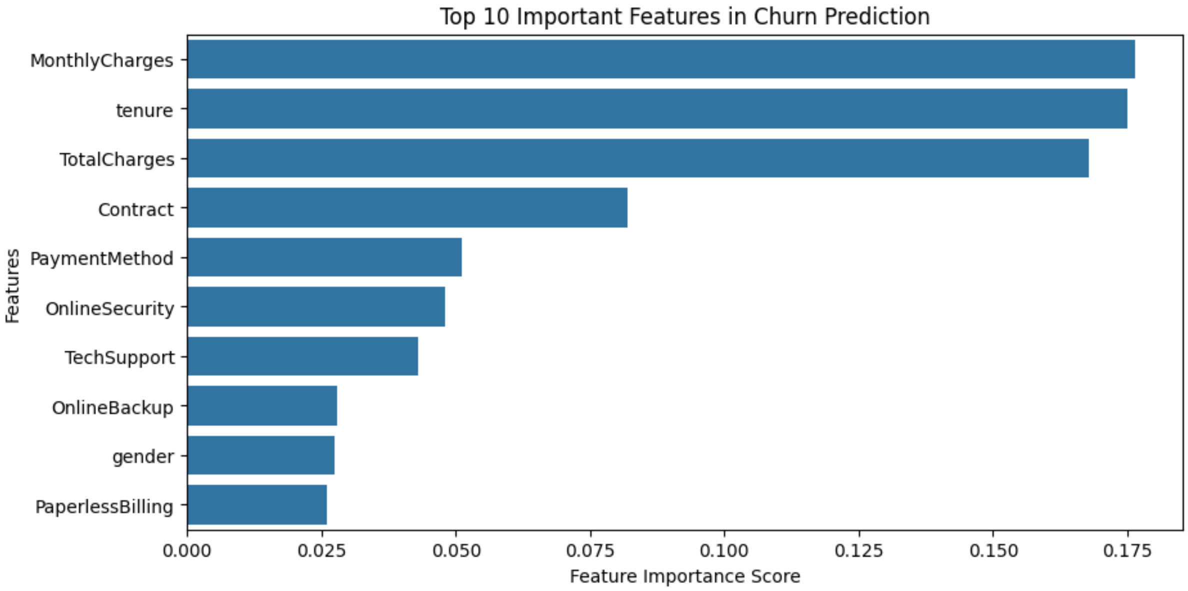
```
Accuracy: 0.7998580553584103


 Classification Report:
               precision    recall  f1-score   support

           0       0.83      0.91      0.87      1036
           1       0.67      0.49      0.56       373

    accuracy                           0.80      1409
   macro avg       0.75      0.70      0.72      1409
weighted avg       0.79      0.80      0.79      1409
```




Top 10 Important Features in Churn Prediction
Features
MonthlyCharges
tenure
TotalCharges
Contract
PaymentMethod
OnlineSecurity
TechSupport
OnlineBackup
gender
PaperlessBilling
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Importance Score