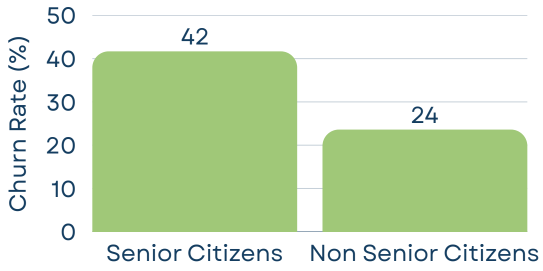
Customer Churn Analysis Report

This report details an analysis of customer churn using a dataset from a telecommunications company. The primary objective of this project is to identify the key factors that contribute to customer churn and to build a predictive model that can accurately identify customers who are likely to churn. The analysis is structured into three main phases: Exploratory Data Analysis (EDA), Data Preprocessing, and Machine Learning Modeling.

Made by
Riyaz Maind
riyazmaind@gmail.com

Exploratory Data Analysis (EDA)

Demographics:



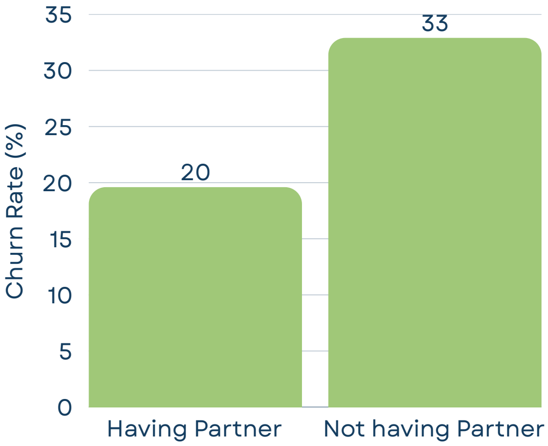
For senior citizens churn rate is HIGH! **41.6%**

The churn rate was found to be fairly consistent across genders, with females having a churn rate of 26.9% and males at 26.1%. However, a significant difference was observed with senior citizens, who have a much higher churn rate of **41.6%** compared to non-senior citizens at 23.6%.

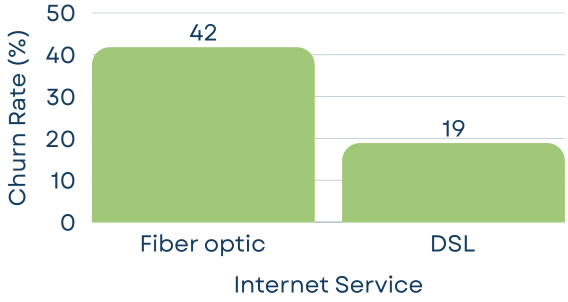
Customer Relationships:

Churn Rate for single people is **32.9%**

Customers without a partner exhibited a higher churn rate of 32.9%, while those with a partner had a rate of 19.6%. Similarly, customers without dependents showed a churn rate of 31%, indicating that single customers or those without family responsibilities are more likely to churn.



Service Usage:

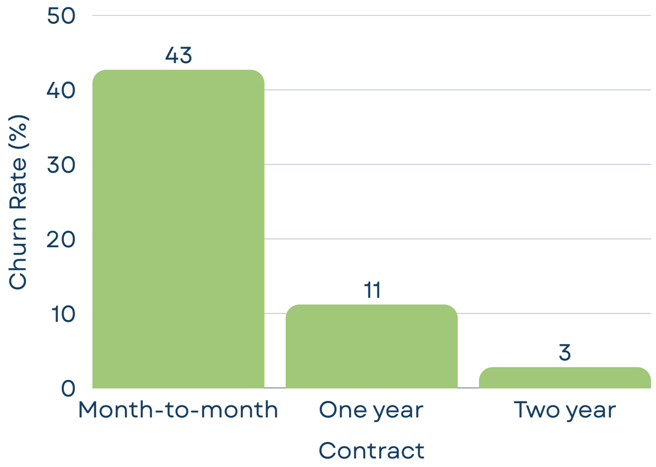


Churn rate for people using Fiber Optics is **41.8%**

The type of internet service strongly influences churn. Customers using Fiber Optic service had the highest churn rate at 41.8%. Additionally, customers who did not have online security, online backup, or tech support services also showed high churn rates of 41.7%, 39.9%, and 41.6% respectively.

Exploratory Data Analysis (EDA)

Contract and Billing:



Churn for people having monthly contract is HIGH! **42.7%**

The length of a customer's contract and their payment method were also significant predictors of churn. Customers on a month-to-month contract had a very high churn rate of 42.7%, while those on one-year or two-year contracts had much lower rates. The electronic check payment method was associated with the highest churn rate at 45.2%.

Tenure and Charges:

Churn was notably high for new customers with less tenure. Furthermore, customers with low total charges and high monthly charges showed a tendency to churn more frequently.

Data Preprocessing

The dataset was prepared for machine learning models through a series of preprocessing steps:

- **Feature Engineering and Cleaning:**

The customerID column was dropped as it holds no predictive value. The TotalCharges column, which was initially of object type and contained some missing values, was converted to a numeric type, and the missing values were imputed using the median of the column.

- **Categorical Encoding:**

Binary categorical columns like gender, Partner, and Churn were converted to numerical representations (0 and 1). The Contract column was also encoded numerically. Other nominal categorical columns like InternetService, PaymentMethod, and various service-related columns were one-hot encoded using pd.get_dummies.

- **Feature Scaling:**

The numerical features of the training and test data were standardized using StandardScaler to ensure that all features contributed equally to the model's performance.

Modeling, Evaluation and Conclusion

Several machine learning models were trained and evaluated on the preprocessed data. The performance was assessed using key metrics such as accuracy and ROC-AUC score, as the latter is particularly useful for imbalanced datasets like this one.

- **Logistic Regression:** Achieved a test accuracy of approximately 80.7%.
- **Random Forest Classifier:** The initial model had a ROC-AUC score of 0.826.
- **XGBoost Classifier:** This model showed a test accuracy of 79.1% and a ROC-AUC score of 0.828.
- **Support Vector Classifier (SVM):** The SVM model's performance was an accuracy of 79.0% and a ROC-AUC score of 0.792.
- **Tuned Random Forest Classifier:** Hyperparameter tuning was performed using GridSearchCV to optimize the Random Forest model's performance. The best parameters found were {'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 300}. This optimized model achieved a test ROC-AUC score of 0.843, which was the highest among all the models. The model also demonstrated a recall of 0.76 for the churn class, indicating its effectiveness in correctly identifying customers at risk of churning.

Conclusion

The Tuned Random Forest classifier is the most effective model for predicting customer churn in this analysis. While other models, such as Logistic Regression and XGBoost, showed high overall accuracy, the Tuned Random Forest's superior ROC-AUC score and strong recall make it the most reliable choice for a business objective focused on proactively retaining at-risk customers. The findings from the EDA section provide actionable insights, suggesting that focusing on customers with short tenure, month-to-month contracts, and those using specific services could be a good starting point for churn mitigation strategies.

GitHub Repo