TELECOM CUSTOMER CHURN ANALYSIS

LEEJO ABY ABRAHAM

29TH JULY 2024

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report presents a comprehensive analysis of customer churn using a dataset from a telecommunications company. The primary objective is to predict customer churn and identify the key factors contributing to it.

Random Forest Classifier was used for its robustness and ability to handle large datasets with higher dimensionality. The analysis revealed that certain features such as monthly charges, contract period and data usage significantly impact customer churn. The report also discusses the model's performance, possible flaws, and recommendations for future work.

INTRODUCTION

Background Information: Customer churn is a critical issue for telecommunications companies as it directly affects revenue and profitability. Understanding the factors that lead to churn can help in developing strategies to retain customers.

Purpose of the Analysis: The main purpose of this analysis is to build a predictive model to identify customers who are likely to churn and understand the key factors influencing their decision.

Scope and Objectives:

- Predict customer churn using machine learning models.
- Identify key features that contribute to customer churn.
- Provide actionable insights to reduce churn rates.

DATA DESCRIPTION

Data Source: The dataset is part of IBM Skill Network Labs.

Dataset Description: The dataset contains 7043 observations with 23 features.

Data Dictionary:

Listed below are the most prominent features of the dataset:

months	Number of months the customer has been with the company
multiple	Indicates if a customer has multiple lines (binary)
gb_mon	Monthly data usage in GBs
security	If the customer has opted for online security add-on (binary)
backup	If the customer has opted for online backup add-on (binary)
protection	If the customer has opted for device protection add-on (binary)
support	If the customer has opted for tech support add-on (binary)
unlimited	If the customer has unlimited data
contract	Type of contract (monthly, annual or bi-annual)
paperless	If the customer is enrolled for paperless billing (binary)
monthly	Monthly charge
satisfaction	Customer satisfaction score
churn_value	*Target variable* – If a customer has churned (binary)

DATA PREPROCESSING

Data Cleaning:

- The dataset does not contain missing values.
- The dataset does not contain duplicate values.
- Outliers were identified and handled appropriately.

Data Transformation:

- Categorical features were encoded using one-hot encoding.
- Numerical features were scaled using min-max scaler.

Feature Selection:

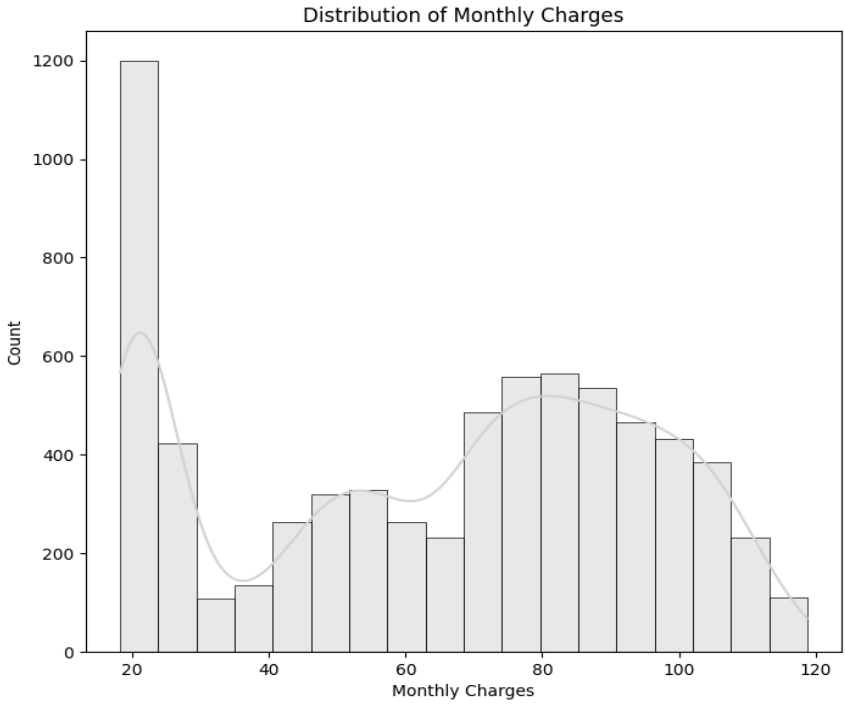
All features were selected for modelling.

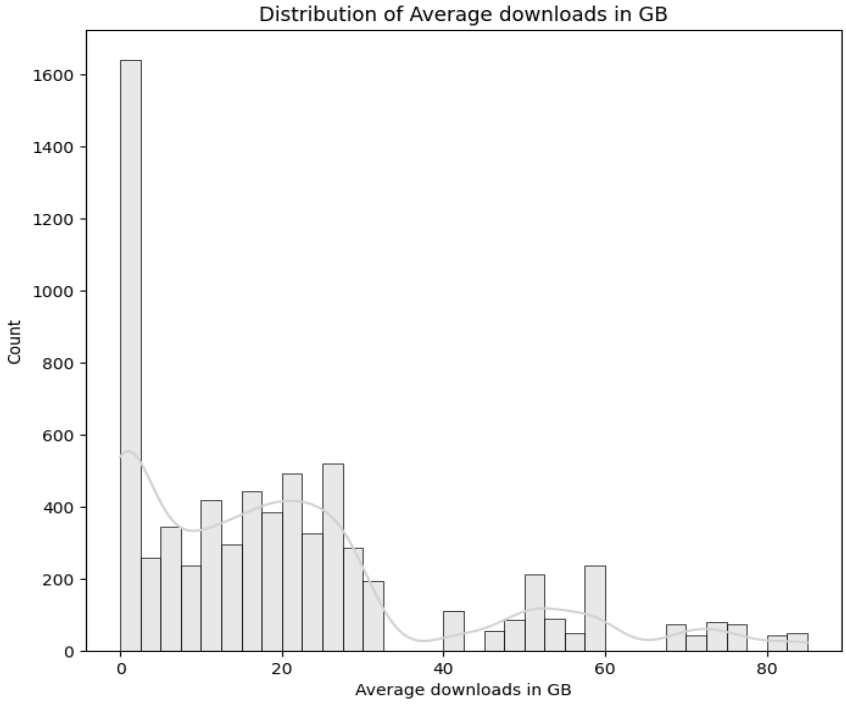
EXPLORATORY DATA ANALYSIS (EDA)

Key Insights from EDA:

- Months (Tenure): Customers with shorter tenure are more likely to churn. The distribution is slightly skewed towards shorter tenure.
- Monthly Charges: Higher monthly charges are associated with an increased likelihood of churn.
- Average Downloads in GB: Most customers have low to no data usage, with a significant peak at 0 GB, and the distribution is highly right skewed indicating few customers with high data usage.

Distribution of Monthly Charges & Average downloads in GB:





MODEL SELECTION AND EVALUATION

Bagging can significantly benefit this dataset by enhancing model stability, minimizing overfitting, and increasing predictive accuracy using ensemble learning. This method is particularly effective for complex datasets with numerous features and potential interactions among them.

Choice of Models:

RandomForestClassifier:

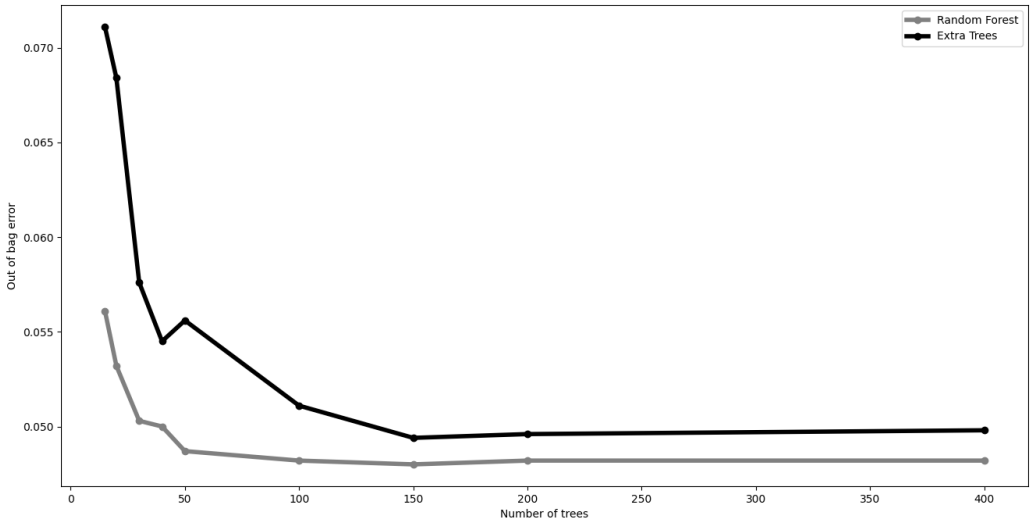
- RandomForestClassifier handles high-dimensional data well and provides feature importance metrics to identify key factors affecting churn.
- It reduces overfitting by averaging multiple decision trees and captures non-linear relationships, making it suitable for complex datasets like this one.

ExtraTreesClassifier:

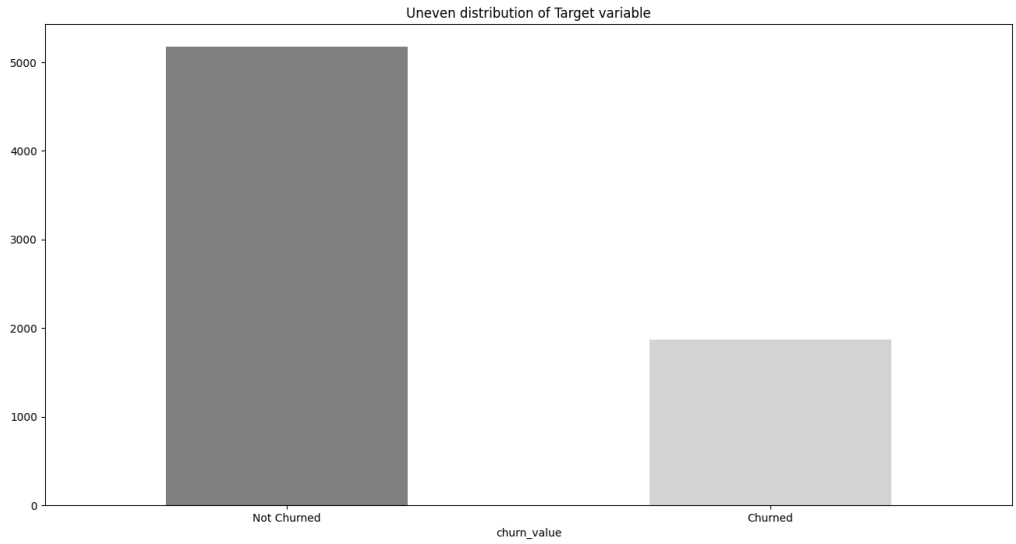
- ExtraTreesClassifier offers better performance and faster computation times by using the entire dataset for each tree.
- It reduces prediction variance for more stable models and shares benefits with RandomForestClassifier, such as handling high-dimensionality and providing feature importance.
- Its randomized splits result in diverse trees and improved generalization.

Out-of-bag error:

Since RandomForestClassifier generated a lower OOB error score, it was chosen as the final model with `n_estimators` as `150`.



Fitting the Model:



StratifiedShuffleSplit was used to split the data into training and testing sets while maintaining the proportion of the target classes, with the test set containing 1500 observations. The training set was fit using the RandomForestClassifier.

Model Evaluation Metrics:

Score attained on train set:

Metrics	Score
Accuracy	1.000
Precision	1.000
Recall	0.997
F1 Score	0.999
AUC	0.999

RESULTS

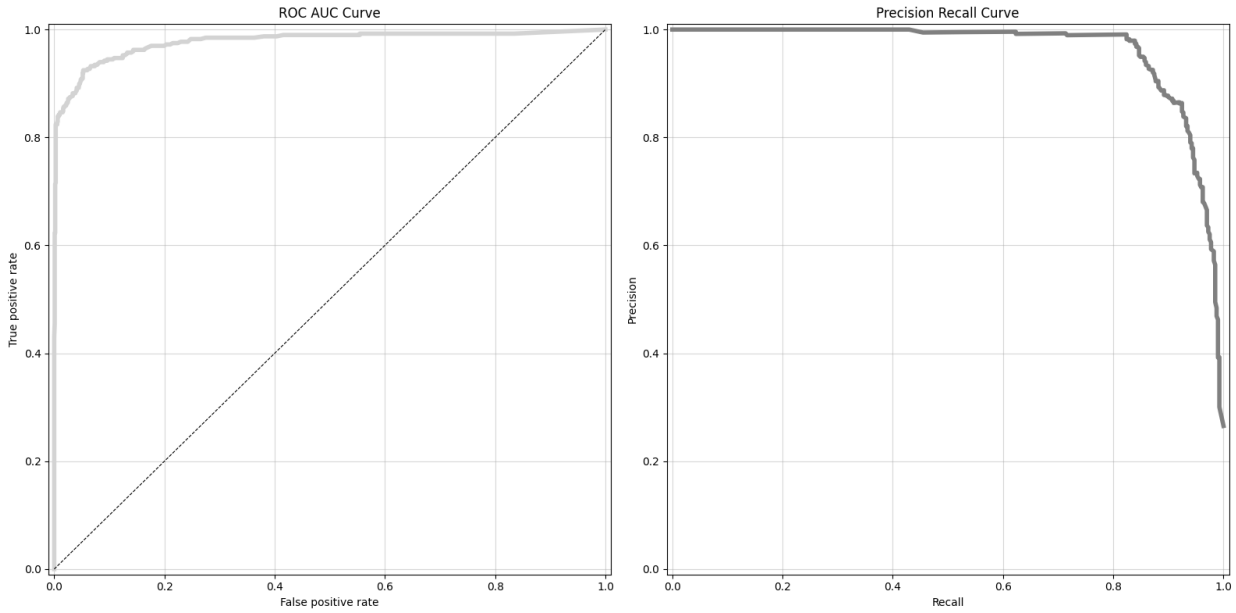
The model evaluation for predicting customer churn was conducted using several performance metrics. The model achieved an accuracy of 94.8%, indicating that the model correctly predicted 94.8% of the instances in the test set. The precision of 93.96% reflects the proportion of true positive churn predictions out of all positive predictions, showcasing the model's ability to minimize false positives effectively. A recall of 85.93% signifies that the model successfully identified 85.93% of the actual churn cases, which is crucial for capturing the at-risk customers.

Moreover, the F1 score, which balances precision and recall, was 89.76%, demonstrating a strong overall performance of the model in predicting churn. The AUC value of 91.97% highlights the model's excellent capability to distinguish between customers who churn and those who do not.

Score attained on test set:

Metrics	Score
Accuracy	0.948
Precision	0.940
Recall	0.859
F1 Score	0.898
AUC	0.920

ROC AUC Curve and Precision Recall Curve



The ROC AUC Curve (left) shows the true positive rate against the false positive rate, with the area under the curve (AUC) being 0.9197. This high AUC value indicates that the model has a strong ability to distinguish between customers who will churn and those who will not.

The Precision-Recall Curve (right) demonstrates the trade-off between precision and recall for different threshold values. The curve shows high precision and recall values, confirming that the model performs well in identifying true positive churns while minimizing false positives.

DISCUSSION

Implications of the Findings:

The findings from the customer churn analysis have significant implications for the telecommunications company. By identifying the key factors that contribute to customer churn, the company can develop targeted strategies to retain customers. The analysis revealed that features such as tenure and monthly charges play a crucial role in predicting churn. Customers with shorter tenure and higher monthly charges are more likely to churn. This insight suggests that the company should focus on improving customer satisfaction and providing incentives for long-term contracts, especially for new customers and those with high monthly charges.

Limitations of the Analysis:

Despite the strong performance of the models, there are several limitations to this analysis. One potential limitation is the risk of overfitting due to the complexity of the RandomForestClassifier algorithm and the limited data available. Although cross-validation and the use of out-of-bag (OOB) scores help mitigate this risk, additional data and further validation are needed to ensure the model's generalizability.

Another limitation is the class imbalance in the dataset, which may affect the model's performance. Although stratified sampling

was used to address this issue, more advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique) could further improve model performance.

Recommendations for Future work:

To build on the findings of this analysis, future work should focus on collecting additional data to enhance the model's accuracy and generalizability. This could include data from different time periods, regions, and customer segments to capture a more comprehensive picture of customer churn.

Additionally, exploring alternative models such as Gradient Boosting and Support Vector Machines could provide further insights and improve predictive performance. Implementing more advanced techniques to address class imbalance, such as SMOTE, could also enhance model performance. Finally, conducting a cost-benefit analysis of different retention strategies based on the model's predictions could help the company allocate resources more effectively and reduce churn rates.

CONCLUSION

This analysis provides valuable insights into customer churn dynamics and offers actionable recommendations for the telecommunications company to mitigate churn. By leveraging the predictive power of machine learning models, the company can proactively identify at-risk customers and implement effective retention strategies, ultimately improving customer loyalty and profitability. The top five features influencing churn were found to be customer satisfaction, tenure, monthly charges, contract type, and monthly data usage. These features should be the focal points of retention strategies to effectively reduce churn rates.

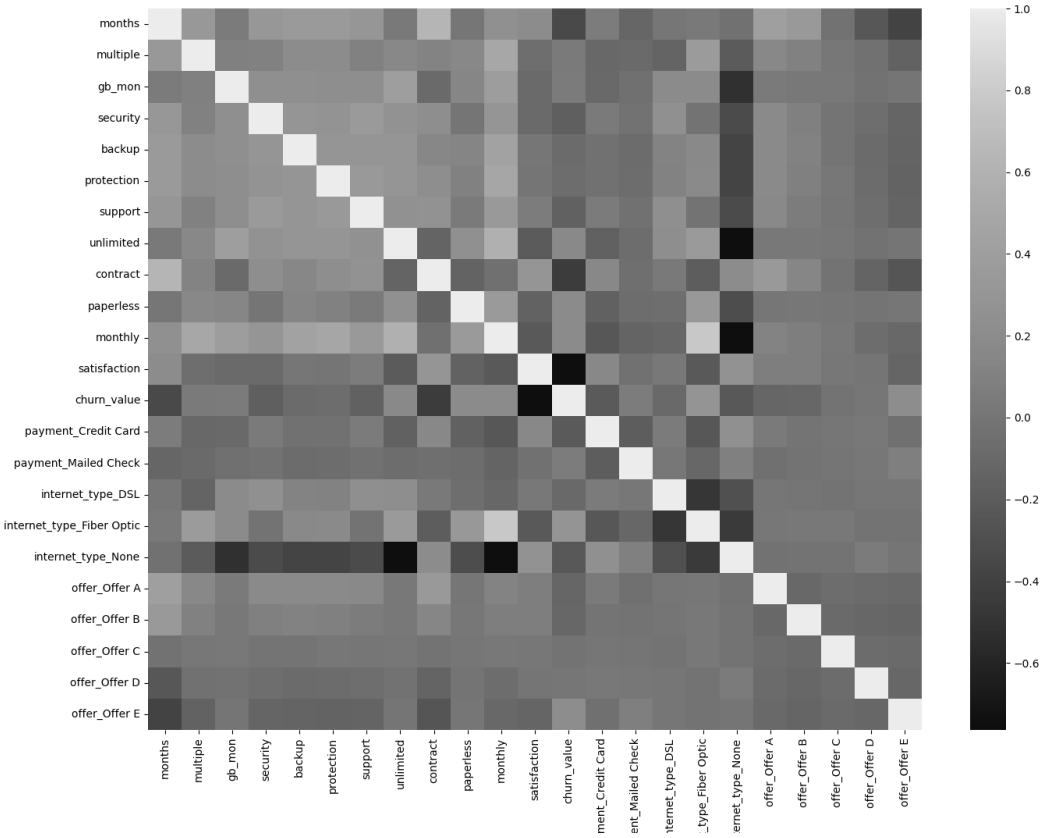
Recommendations:

Based on the insights gained from this analysis, several recommendations can be made to reduce customer churn:

- Targeted Retention Strategies: Focus on retaining customers with shorter tenure and higher monthly charges by offering incentives, personalized services, and loyalty programs.
- Improve Customer Satisfaction: Enhance customer support and service quality, particularly for customers with month-to-month contracts, to increase satisfaction and reduce churn.
- Long-Term Contracts: Encourage customers to switch to longer-term contracts by providing attractive offers and discounts, as longer contracts are associated with lower churn rates.

APPENDECES

Correlation Heatmap between variables:



This correlation matrix is a valuable tool for understanding the relationships between different features and guiding the feature selection process in the analysis.

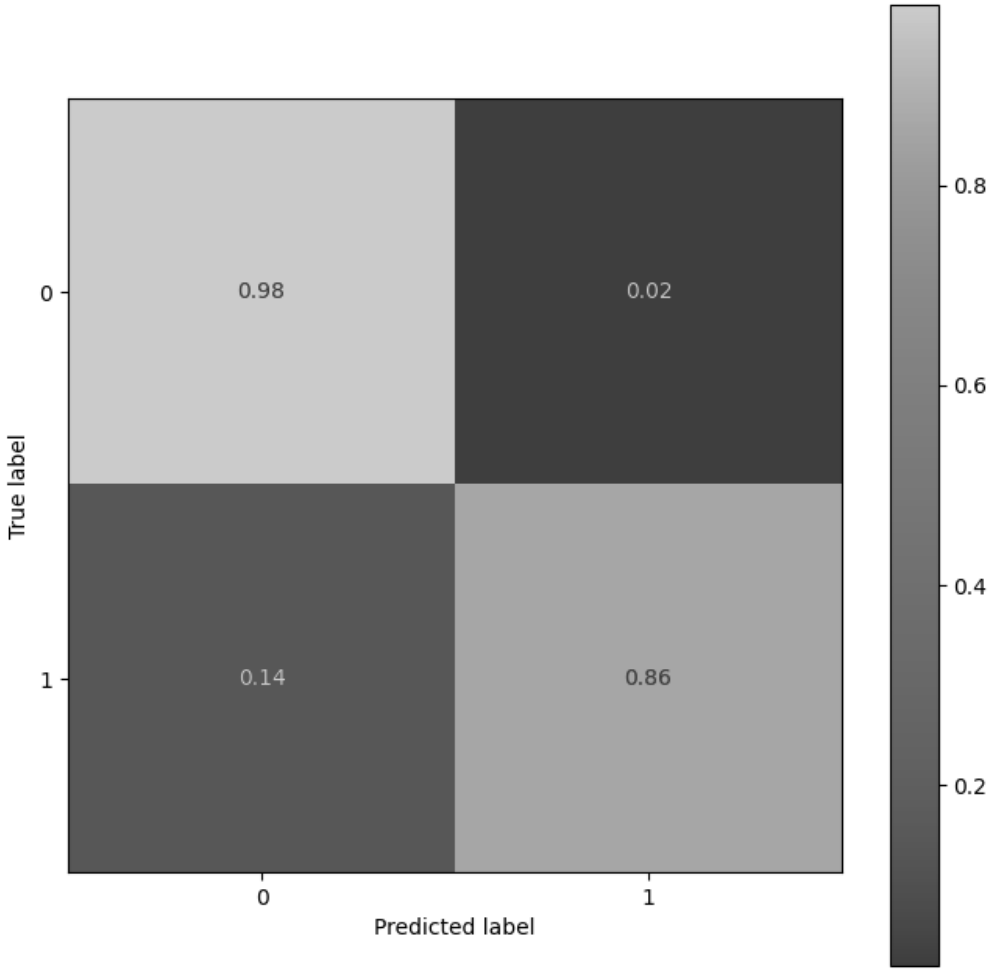
Key Observations:

- Tenure and Monthly Charges: There is a moderate positive correlation between tenure and monthly charges, indicating

that customers who have been with the company longer tend to have higher monthly charges.

- Contract Type and Churn: There is a noticeable negative correlation between contract type and churn, suggesting that customers with longer contracts are less likely to churn.
- Customer Satisfaction: This feature shows significant correlations with several other variables, highlighting its importance in predicting customer churn.

Confusion Matrix of test scores:



This confusion matrix complements the evaluation metrics discussed earlier, providing a comprehensive understanding of the model's performance in predicting customer churn.

Key Insights:

- The high true positive rate (86%) and true negative rate (98%) demonstrate the model's effectiveness in correctly identifying both churn and non-churn cases.
- The low false positive rate (2%) and moderate false negative rate (14%) indicate that while the model is very good at minimizing incorrect churn predictions, there is still room for improvement in capturing all actual churn cases.