Assignment 2

Classification Models

Shashikanth Senthil Kumar
Student ID: 25218722

2024 SPRING

36106 - Machine Learning Algorithms and Applications
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

1. Business Understanding **2**

a. Business Use Cases 2

b. Key Objectives 3

2. Data Understanding **4**

a. Dataset overview 4

b. Variables/Features and Their Relevance 4

c. Exploratory Data Analysis (EDA) Insights 5

d. Data Limitations 5

3. Data Preparation **6**

a. Data Cleaning 6

b. Feature Selection 7

c. Feature Engineering 7

d. Data Preprocessing 8

4. Modeling **9**

a. Machine Learning Algorithm Used 9

b. Rationale Behind Algorithm Selection 9

c. Hyperparameter Tuning 9

d. Model Selection Criteria 10

5. Evaluation **11**

a. Results and Analysis 11

b. Business Impact and Benefits 13

c. Data Privacy and Ethical Concerns 14

6. Conclusion **15**

7. References **16**

8. Appendices **17**

1. Business Understanding

a. Business Use Cases

The project aims to predict customer churn within the next month, allowing the company to implement targeted retention strategies.

- **Customer Retention Management**: Identify high-risk customers to offer 50% discounts for three months, improving customer loyalty and reducing revenue loss.
- **Marketing Optimization**: Focus marketing efforts on at-risk customers with personalized offers, increasing retention efficiency.
- **Subscription Renewal Strategy**: Use churn predictions to prioritize discounts for customers most likely to renew, ensuring cost-effective retention.

Challenges:

- **Imbalanced Data**: Churn is a minority class, making it difficult to train models that balance false positives and negatives.
- **Customer Behavior Complexity**: Multiple factors such as usage patterns and payment methods complicate feature selection and model performance.

Opportunities:

- **Cost-Efficient Retention**: Targeting the right customers with discounts minimizes overall retention costs while maximizing retention success.
- **Personalized Engagement**: Customized offers based on predictive insights enhance customer interactions and reduce churn.
- **Improved Lifetime Value**: Retaining customers for longer periods increases the company's overall lifetime value per customer.

b. Key Objectives

- **Predict Customer Churn Accurately**: Train models like Logistic Regression, Decision Trees, Random Forest, and XGBoost Classifiers to predict churn.
- **Optimize Retention Strategies**: Use churn predictions to target high-risk customers with retention offers.
- **Evaluate Model Performance**: Use metrics such as Recall, Precision, and F1-Score to select the best model.
- **Provide Actionable Insights**: Identify churn drivers and recommend data-driven strategies for retention and marketing teams.

Stakeholders and Their Requirements:

- **Customer Retention Team**: Focuses on identifying at-risk customers and designing targeted campaigns to retain them, using churn predictions to prioritize efforts.
- **Marketing Team**: Requires accurate segmentation to create personalized offers that will encourage at-risk customers to stay, based on insights from the model.
- **Financial Team**: Needs to assess the impact of churn on revenue and ensure the discount strategy is profitable, using churn forecasts to guide decision-making.
- **Management**: Seeks to reduce churn and maintain growth while ensuring customer satisfaction, using model insights to refine retention strategies.

Project Goals:

- **Model Training**: Train and tune various machine learning models for high churn prediction accuracy.
- **Performance Assessment**: Evaluate models using relevant performance metrics like Recall to ensure effective predictions.
- **Business Impact**: Implement churn predictions to improve retention, reduce churn, and boost revenue.
- **Recommendations**: Provide actionable insights for marketing and retention efforts based on model results.

2. Data Understanding

a. Dataset Overview

The dataset used for this project is aimed at predicting customer churn in a subscription-based service. The final dataset used for this project, derived from 10 separate CSV files that were merged to create a unified dataset with 34,586 rows and 31 columns. The data includes a mix of numerical and categorical features related to the customers.

Data Collection and Sources:

The dataset has been collated by the Marketing team. It appears to be collected from multiple sources related to customer interactions, subscription details, and demographic information. However, the specific sources of the data or the methods used for data collection are not provided in the brief. This lack of information on data provenance might affect the interpretability and reliability of the model's predictions.

b. Variables/Features and Their Relevance

Key features present in the dataset include:

- **SubscriptionType**: Categorizes the type of subscription (standard, basic, premium), which may significantly influence churn risk.
- **PaymentMethod**: Different payment methods (credit card, electronic check, etc.) may correlate with customer retention rates.
- **MonthlyCharges**: Higher monthly charges may indicate a greater risk of churn, especially if customers feel they are not receiving adequate value.
- **AccountAge**: Older accounts may be less likely to churn due to established customer relationships, showing a strong relationship with churn prediction.
- **SupportTicketsPerMonth**: Higher support ticket volume may indicate customer dissatisfaction, potentially increasing churn risk.

c. Exploratory Data Analysis (EDA) Insights

- **Target Variable (Churn)**:
  - The primary objective is to predict customer churn, a binary variable where '1' indicates churn and '0' indicates no churn.
  - Approximately 17.75% of customers have churned, while 82.25% have not (See fig-1), indicating a class imbalance that could influence model performance.
- **Monthly Charges**
  - Higher Monthly Charges correlate with increased churn rates (See fig-4), indicating potential dissatisfaction with value .
  - Retention strategies for high-paying customers, such as enhanced features, are recommended.
- **Account Age**
  - Newer customers (under 50 months) are at higher churn risk (See fig-2), while long-term customers tend to be more loyal.
  - Churned customers typically have a shorter Account Age, highlighting the need for early engagement. Personalized support for newer customers could help improve retention.
- **Subscription Type**
  - Standard is the most common subscription type (See fig-3), analyze churn rates across all types for trends.
  - High churn in Premium subscriptions may signal a disconnect in perceived value.

d. Data Limitations

Several limitations were identified during the data exploration process:

Irrelevant Features: The dataset may include irrelevant personal identifiers, which should be removed to avoid privacy concerns and prevent data leakage.

Data Discrepancy: Variations in customer behavior across different segments might impact the model's generalizability, particularly for underrepresented groups.

Class Imbalance: The imbalance between churn and non-churn customers can bias predictions. Model training must address this through sampling strategies or by using appropriate evaluation metrics such as F1-score or precision.

Despite these limitations, the dataset offers valuable features related to customer behavior, subscription details, and support interactions.

3. Data Preparation

The data preparation process for modeling involved several key steps aimed at ensuring the dataset was clean, relevant, and suitable for analysis. Below are the detailed actions taken in terms of data cleaning, preprocessing, and feature engineering.

a. Data Cleaning:

i. **Dataset Copying:**

- **Step Taken:** Datasets were copied to ensure that the original data remained unchanged throughout the cleaning process.
- **Purpose:** Preserves the integrity of the original data and allows for safe experimentation during cleaning.

ii. **Handling Missing Values:**

- **Step Taken:** The datasets were checked for missing values, but none were found.
- **Purpose:** Ensures data integrity and avoids issues that missing values could cause during modeling.

iii. **Standardizing Data :**

- **Step Taken:** Categorical data entries were standardized by converting all string entries to lowercase using a loop across all columns containing object types.
- **Purpose:** Ensures uniformity in categorical text data, enhancing accuracy and reliability for analysis and modeling.

iv. **Removing Duplicated Values:**

- **Step Taken:** Checked for and confirmed that there were no duplicated values in any dataset.
- **Purpose:** Ensures that each data point is unique and prevents bias in the analysis.

b. Feature Selection

i. **Removal of Personal Identifiable Information (PII):**

- **Step Taken:** Personal details such as names, addresses, and other identifiable information (e.g., 'FirstName', 'LastName', 'StreetName', 'Postcode', 'Ethnicity', 'Gender', 'CustomerID') were removed from the dataset.
- **Purpose:** Safeguards customer privacy and eliminates features that do not contribute to predicting churn, reducing potential bias.

ii. **Correlation Analysis:**

- **Step Taken:** Conducted correlation analysis to assess the relationship between numerical features and churn, identifying impactful features and detecting multicollinearity (See fig-5). Multicollinearity between 'TotalCharges' and 'MonthlyCharges' indicated a potential need to drop one of these features. Notable correlations included modest positive correlations for 'MonthlyCharges' and 'SupportTicketsPerMonth', and negative correlations for 'AccountAge' and 'TotalCharges'.
- **Purpose:** Focuses on retaining features most likely to influence churn outcomes, improving model performance and interpretability.

c. Feature Engineering

i. **Creating Interaction Terms:**

- **Step Taken:** Introduced new features to capture the combined effects of existing variables. For instance, 'UserRating_SupportTicketsInteraction' was created by multiplying 'UserRating' with 'SupportTicketsPerMonth' to evaluate if low ratings combined with high support interactions indicate a higher churn risk.
- **Purpose:** Enhances the model's ability to detect complex relationships between features, potentially revealing hidden patterns that influence churn risk.

ii. **Categorizing Continuous Variables:**

- **Step Taken:** Developed a new categorical feature, 'MonthlyChargeTier', by binning 'MonthlyCharges' into tiers (e.g. Very Low, Low, Medium, High) to analyze price sensitivity related to churn.
- **Purpose:** Simplifies the relationship between monthly charges and churn, making it easier to interpret and analyze how different pricing tiers affect customer retention

d. Data Preprocessing

i. **Split Datasets**

- **Step Taken:** Data Splits into 80% training, 10% validation, and 10% testing.
- **Purpose:** Ensures proper training, hyperparameter tuning, and unbiased final evaluation.

ii. **Encoding Ordinal Categorical Variables**

- **Step Taken:** Applied OrdinalEncoder to 'SubscriptionType', 'PaymentMethod', and 'MonthlyChargeTier'.
- **Purpose:** Preserves the order of categories, enabling the model to interpret ordinal relationships.

iii. **Label Encoding Binary Categorical Features**

- **Step Taken:** Used LabelEncoder for 'PaperlessBilling', 'MultiDeviceAccess', 'ParentalControl', and 'SubtitlesEnabled'.
- **Purpose:** Converts binary features into 0 and 1 for model compatibility.

iv. **Encoding Nominal Categorical Features**

- **Step Taken:** Used OrdinalEncoder for 'DeviceRegistered', 'ContentType', and 'GenrePreference'.
- **Purpose:** Converts nominal features into numerical values for machine learning algorithms.

4. Modeling

a. Machine Learning Algorithms Used:

- **Logistic Regression**: Predicts the likelihood of customer churn based on a linear relationship between features and the log-odds of churn. Simple and interpretable baseline model.
- **Decision Tree**: Splits data into branches based on feature values to predict churn. It's easy to interpret but can overfit without proper pruning.
- **Random Forest**: An ensemble of decision trees that reduces overfitting and improves accuracy by averaging predictions from multiple trees built on random data subsets.
- **XGBoost**: A high-performance gradient boosting model that builds trees sequentially to correct previous errors. It's fast, handles complex interactions, and is ideal for churn prediction with imbalanced datasets.

b. Rationale Behind Algorithm Selection

- **Logistic Regression**: Simple, interpretable baseline for churn prediction.
- **Decision Tree**: Captures non-linear relationships and key churn drivers, though prone to overfitting.
- **Random Forest**: Reduces overfitting and captures complex patterns using ensemble learning.
- **XGBoost**: High accuracy for imbalanced data, with gradient boosting for error correction and performance optimization.

c. Hyperparameter Tuning

The tuning process aims to optimize model performance by selecting the best combination of hyperparameters.Although grid search was initially planned to optimize we decided to manually using itertools

Logistic Regression:

- **'C'**: Regularization strength; lower values impose more regularization.
- **'solver'**: Algorithm for optimization (liblinear for small datasets, lbfgs/saga for larger ones).
- **'max_iter'**: Number of iterations for convergence.
- **'class_weight'**: Assigns weights to handle class imbalance.

Decision Tree:

- **'criterion'**: Measures split quality (gini or entropy).
- **'max_depth'**: Limits tree depth to prevent overfitting.
- **'min_samples_split'**: Minimum samples required to split a node.
- **'min_samples_leaf'**: Minimum samples in a leaf node.
- **'class_weight'**: Handles class imbalance by adjusting weights.

Random Forest:

- **'n_estimators'**: Number of trees in the forest.
- **'max_depth'**: Limits tree depth for each estimator.
- **'min_samples_split'**: Controls node splitting, impacting model complexity.
- **'min_samples_leaf'**: Ensures minimum samples in leaf nodes to reduce overfitting.
- **'class_weight'**: Adjusts for class imbalance.

XGBoost:

- **'n_estimators'**: Number of boosting rounds.
- **'max_depth'**: Depth of trees for capturing complex patterns.
- **'learning_rate'**: Shrinks contribution of each tree to avoid overfitting.
- **'subsample'**: Fraction of data used for fitting trees.
- **'scale_pos_weight'**: Adjusts weights to account for class imbalance.

d. Model Selection Criteria

- **Evaluation Metric**: The performance of each model was assessed using **Recall**, **Precision**, and **F1-Score**, which are critical for balancing the need to identify true churners while minimizing unnecessary discounts. A high recall ensures most at-risk customers are detected, while high precision reduces false positives, preventing wasted resources.
- **Model Performance**: Models were evaluated based on these metrics across training, validation, and test datasets. This helped determine which models generalized well to unseen data, balancing between identifying actual churners and avoiding false alarms.
- **Final Selection**: The final model was chosen based on the highest F1-Score, which balances recall and precision, ensuring the best trade-off between catching actual churners and avoiding unnecessary discounts.

5. Evaluation

a. Results and Analysis

Experiment 1: Logistic Regression on Churn Prediction

Hypothesis: Customers with higher monthly charges are more likely to churn.

Results:

- **Best Parameters**:
  C: 0.1, Solver: liblinear, Class Weight: balanced, Max Iterations: 500
- **Performance** :
  - **Training**: Precision: 0.3167, Recall: 0.6920, F1-Score: 0.4346
  - **Validation**: Precision: 0.3128, Recall: 0.6710, F1-Score: 0.4267
  - **Test**: Precision: 0.3079, Recall: 0.6906, F1-Score: 0.4259 (See fig-7)
- **Analysis**:
  The model achieved high recall, successfully identifying a majority of customers likely to churn(See fig-8). However, the lower precision indicates a significant number of false positives, suggesting that some non-churners are misclassified as at risk.

Experiment 2: Decision Tree on Churn Prediction

Hypothesis: The Decision Tree model will identify key drivers of customer churn, with AccountAge and MonthlyCharges having the highest impact on predicting churn.

Results:

- **Best Parameters**:
  Criterion: gini,Max Depth: 5,Min Samples Split: 10,Min Samples Leaf: 4,Class Weight: balanced.

- **Performance Summary**:
  - **Training**: Precision: 0.2861, Recall: 0.7414, F1-Score: 0.4129
  - **Validation**: Precision: 0.2734, Recall: 0.7036, F1-Score: 0.3938
  - **Test**: Precision: 0.2636, Recall: 0.6873, F1-Score: 0.3810 (See fig-9)
- **Analysis**

  The Decision Tree model achieved a high recall, effectively identifying a significant number of churners (See fig-10). However, low precision indicates a considerable number of false positives.

Experiment 3: Random Forest Classifier on Churn Prediction

Hypothesis: The Random Forest Classifier will effectively identify key drivers of customer churn, with ViewingHoursPerWeek and AverageViewingDuration having the highest impact on predicting churn.

Results:

- **Best Parameters:**
  n_estimators: 200,max_depth: 5,min_samples_split: 2,min_samples_leaf: 1,class_weight: balanced.
- **Performance Summary:**
  - **Training:**Precision: 0.3260,Recall: 0.7097,F1-Score: 0.4468.
  - **Validation:**Precision: 0.3037,Recall: 0.6564,F1-Score: 0.4152.
  - **Test:**Precision: 0.3040,Recall: 0.6629,F1-Score: 0.4168 (See fig-11).
- **Analysis:**
  The Random Forest Classifier achieved a high recall but lower precision indicates that there are still false positives (See fig-12). Feature importance analysis revealed that AccountAge, ViewingHoursPerWeek, and AverageViewingDuration are major factors driving churn.

Experiment 4: XGBoost Classifier on Churn Prediction

Hypothesis: The XGBoost classifier will effectively identify key drivers of customer churn, with features like MonthlyCharges, and SubscriptionType having the most significant impact on predicting churn.

Results:

- **Best Parameters:**
  n_estimators: 200,max_depth: 3,learning_rate: 0.1,subsample: 0.8,scale_pos_weight: 3.
- **Performance:**
  - **Training:** Precision: 0.4149, Recall: 0.5672, F1-Score: 0.4792
  - **Validation:** Precision: 0.3721, Recall: 0.4951, F1-Score: 0.4249
  - **Test:** Precision: 0.3635, Recall: 0.4902, F1-Score: 0.4175 (See fig-13)
- **Analysis:**
  High recall but lower precision; important features include AccountAge, AverageViewingDuration, and SupportTicketsPerMonth (See fig-14). Identifying churn drivers allows targeted retention strategies, potentially reducing churn rates.

Comparison:

Logistic Regression: Achieves high recall, effectively identifying many customers likely to churn. However, its lower precision indicates a significant number of false positives, suggesting some non-churners are misclassified as at risk.

Decision Tree: Successfully identifies key drivers of customer churn, such as AccountAge and MonthlyCharges, with high recall. Yet, its low precision points to a considerable number of false positives, indicating misclassification of non-churners.

Random Forest Classifier: Identifies important churn factors, including ViewingHoursPerWeek and AverageViewingDuration, and shows high recall. However, it also suffers from low precision, leading to notable false positives in predictions.

XGBoost Classifier: Demonstrates high recall, indicating effectiveness in identifying at-risk customers. Key features influencing churn include AccountAge and AverageViewingDuration. Its lower precision, though, suggests room for improvement in reducing false positives.

Best Model:

XGBoost Classifier is the best model among those evaluated. It achieves the highest F1-Score, balancing recall and precision effectively compared to the other models. Although improvements in precision are needed, it provides robust predictions for churn identification.

Reason for Selection:

- **XGBoost Classifier** shows the best trade-off between recall and precision, essential for detecting true churners while minimizing unnecessary discounts.
- It effectively identifies key drivers of churn, allowing for targeted retention strategies.
- While it requires optimization to improve precision, its overall performance makes it the most suitable model for this dataset and problem.

In conclusion, the XGBoost Classifier is recommended for its superior ability to predict customer churn while highlighting important influencing factors, ultimately supporting effective churn reduction strategies.

b. Business Impact and Benefits

Important Features: Features such as AccountAge, AverageViewingDuration, and SupportTicketsPerMonth significantly contribute to accurately predicting customer churn. This indicates that businesses should focus on enhancing customer engagement and support to improve retention rates.

Less Important Features : Features like MonthlyCharges and SubscriptionType have minimal impact on churn predictions, suggesting that businesses may not benefit significantly from adjustments in pricing strategies.

Business Use Case: Companies should leverage predictive analytics to identify customers likely to churn and offer targeted discounts. This proactive approach can enhance customer retention and increase revenue through improved subscription renewals.

c. Data Privacy and Ethical Concerns

Data Privacy Implications:

- **Sensitive Information**: The dataset contains personal identifiers such as names, addresses, customer IDs, and demographic details like ethnicity and gender. This sensitive information requires strict handling to safeguard individual privacy.

Steps Taken for Data Privacy:

- **Sensitive Data Removal**: During the feature selection process, personal identifiable information (PII), including names, addresses, customer IDs, and the 'Cohort' feature, was dropped to ensure that no personal data is used in model training.
- **Anonymization**: All identifiable personal information (e.g., first name, last name, and address) was excluded from the analysis to maintain data anonymity and protect customer privacy.

Ethical Considerations:

- **Responsible Data Use**: We ensured that data collection, usage, and model deployment adhered to ethical standards, avoiding any form of discrimination or misuse of data.
- **Bias Mitigation**: To prevent biased predictions, sensitive features such as gender and ethnicity were removed from the dataset during the preprocessing phase, thus aiming for equitable treatment of all customers.

Potential Negative Impacts:

- **Indigenous Communities**: Care was taken to ensure that the model does not perpetuate biases or inequalities, particularly towards Indigenous communities. By addressing these concerns, we aim to develop fair and responsible outcomes in model predictions.

6. Conclusion

In this project, we developed a predictive model to estimate customer churn using features such as SubscriptionType, PaymentMethod, MonthlyCharges, AccountAge, and SupportTicketsPerMonth. By applying algorithms like Logistic Regression, Decision Trees, Random Forest, and XGBoost, we evaluated their effectiveness using Recall, Precision, and F1-Score.

Key findings include:

- The **XGBoost Classifier** outperformed other models, achieving the highest F1-Score and effectively balancing recall and precision.
- Significant predictors of churn included **AccountAge, AverageViewingDuration,** and **SupportTicketsPerMonth**, indicating areas for targeted retention strategies.
- Higher **MonthlyCharges** were associated with increased churn rates, emphasizing the importance of perceived value in retention efforts.

This project successfully created a reliable model for predicting customer churn, providing valuable insights for stakeholders to optimize retention strategies and enhance customer engagement.

Future Work and Recommendations

- Incorporating additional features like customer feedback or behavioral patterns to improve predictions.
- Exploring advanced techniques such as ensemble methods and neural networks for better performance.
- Implementing a real-time monitoring system to adapt to new customer data and refine retention strategies.These steps will help maximize customer loyalty and drive revenue growth.

7. References

King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. https://doi.org/10.1145/2939672.2939785

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, 51-56. https://doi.org/10.25080/Majora-92bf1922-00a

Scikit-Learn: Machine Learning in Python. (2023). *Scikit-Learn Documentation*. Retrieved from https://scikit-learn.org/

Josue Obregon, Aekyung Kim, Jae-Yoon Jung.(2019). RuleCOSI: Combination and simplification of production rules from boosted decision trees for imbalanced classification,Expert Systems with Applications,Volume 126,Pages 64-82,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.02.012

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. CRC Press. https://doi.org/10.1201/9781315139470

8. Appendices

Fig- 1 (Proportion of Churn)

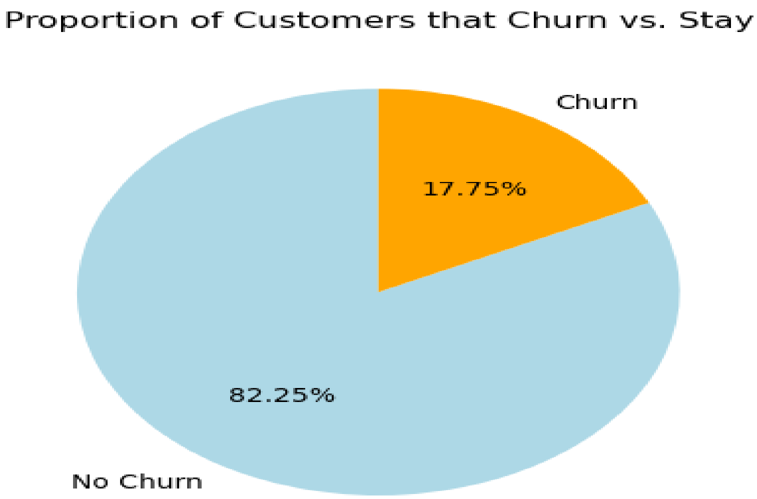


Fig- 2 (Distribution of AccountAge)

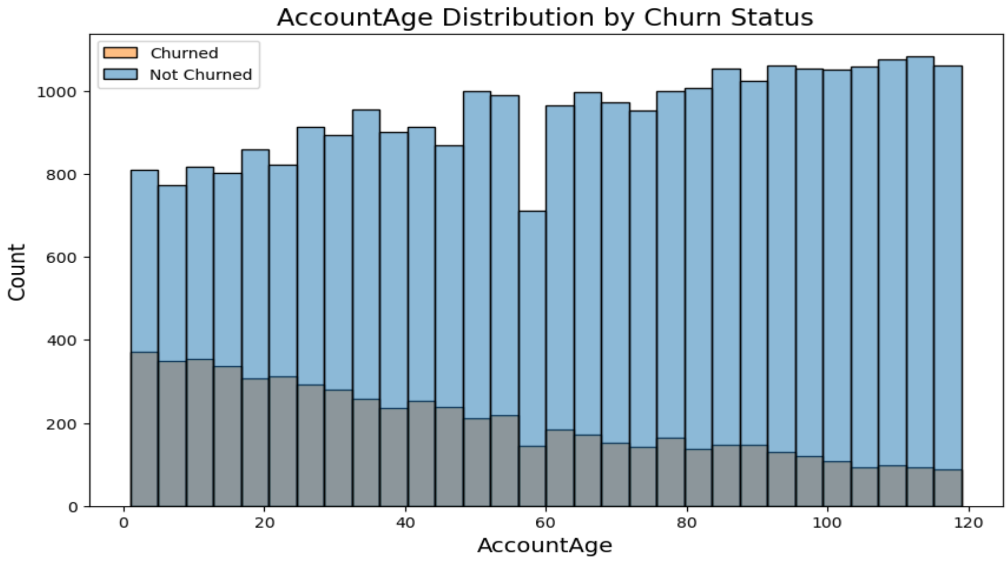


Fig- 3 (Proportion of Subscripion Types)

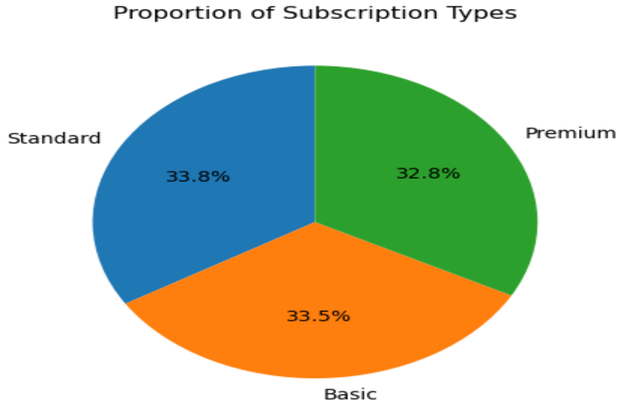


Fig- 4 (Distribution of MonthlyCharges)

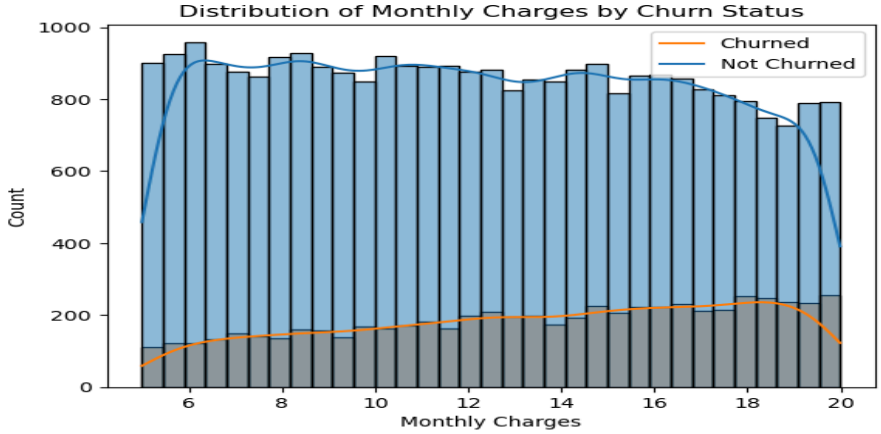
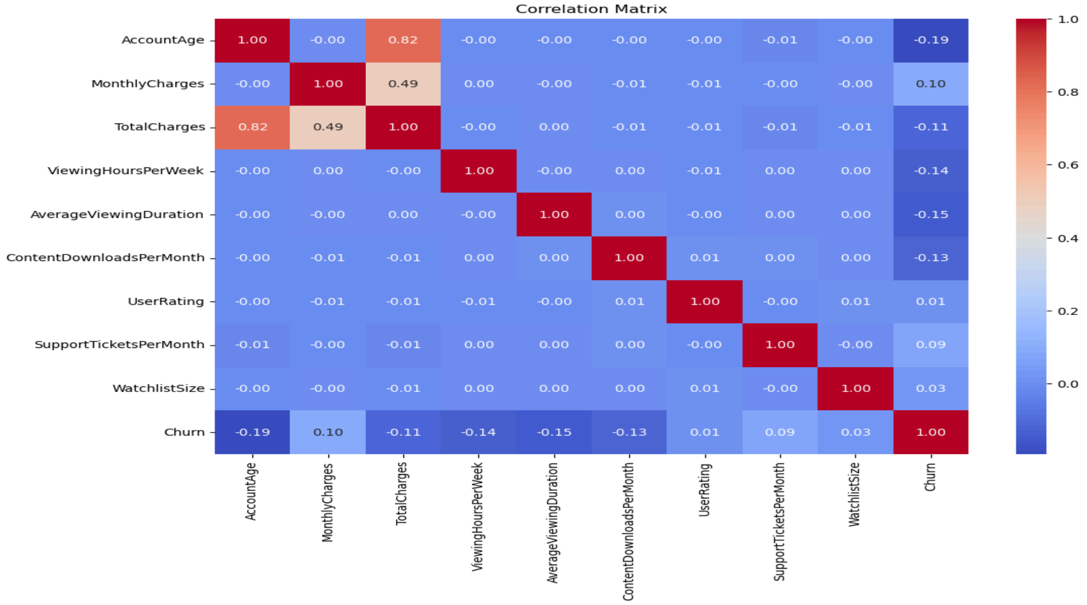


Fig- 5 (Correlation Analysis)



	AccountAge	MonthlyCharges	TotalCharges	ViewingHoursPerWeek	AverageViewingDuration	ContentDownloadsPerMonth	UserRating	SupportTicketsPerMonth	WatchlistSize	Churn
AccountAge	1.00	-0.00	0.82	-0.00	-0.00	-0.00	-0.00	-0.01	-0.00	-0.19
MonthlyCharges	-0.00	1.00	0.49	0.00	-0.00	-0.01	-0.01	-0.00	-0.00	0.10
TotalCharges	0.82	0.49	1.00	-0.00	0.00	-0.01	-0.01	-0.01	-0.01	-0.11
ViewingHoursPerWeek	-0.00	0.00	-0.00	1.00	-0.00	0.00	-0.01	0.00	0.00	-0.14
AverageViewingDuration	-0.00	-0.00	0.00	-0.00	1.00	0.00	-0.00	0.00	0.00	-0.15
ContentDownloadsPerMonth	-0.00	-0.01	-0.01	0.00	0.00	1.00	0.01	0.00	0.00	-0.13
UserRating	-0.00	-0.01	-0.01	-0.01	-0.00	0.01	1.00	-0.00	0.01	0.01
SupportTicketsPerMonth	-0.01	-0.00	-0.01	0.00	0.00	0.00	-0.00	1.00	-0.00	0.09
WatchlistSize	-0.00	-0.00	-0.01	0.00	0.00	0.00	0.01	-0.00	1.00	0.03
Churn	-0.19	0.10	-0.11	-0.14	-0.15	-0.13	0.01	0.09	0.03	1.00

Fig- 6 (Best Baseline model results)

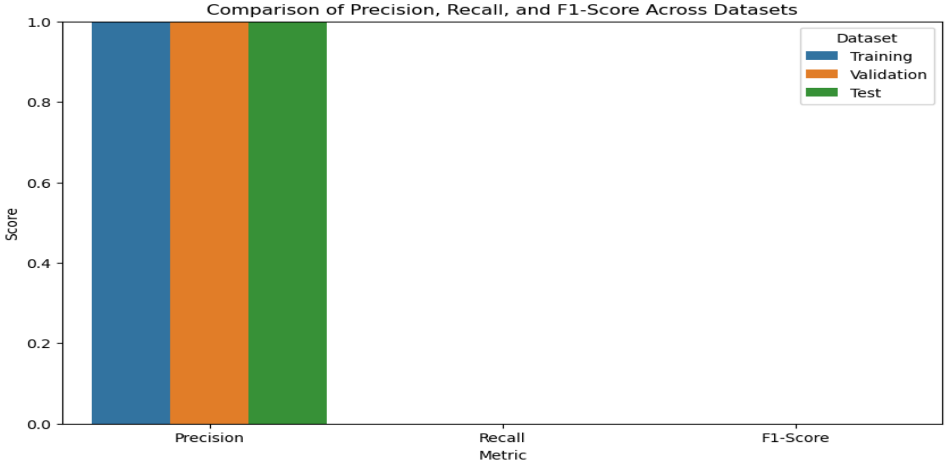


Fig- 7 (Best Logistic Regression model results)

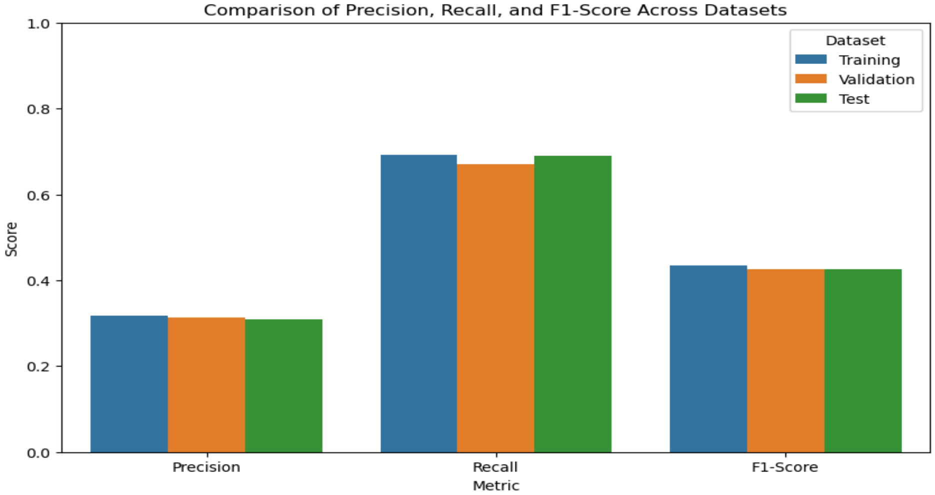


Fig- 8 (Feature Importance for Logistic Regression)

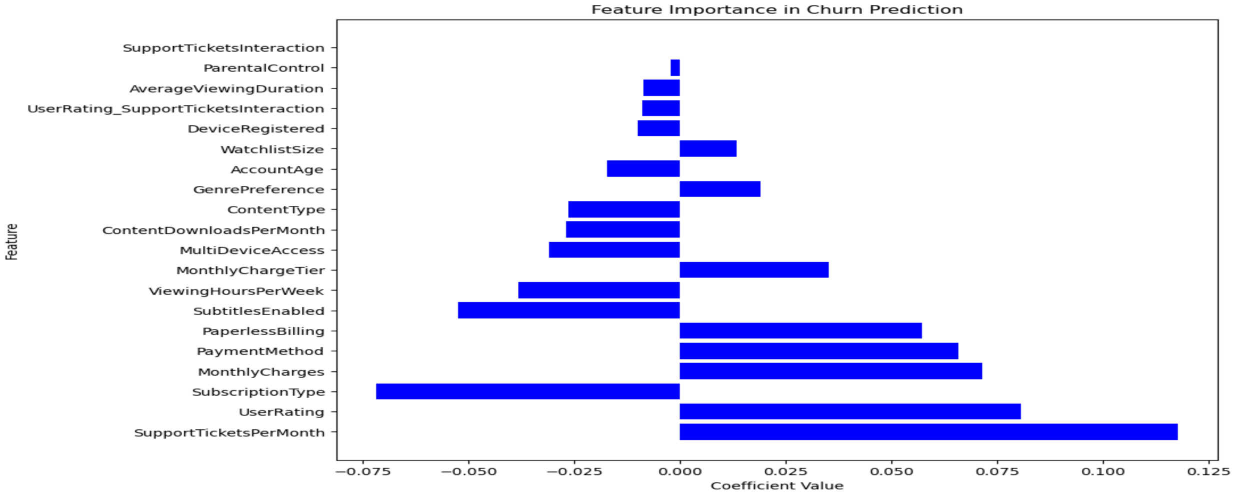


Fig- 9 (Best Decision Tree model results)

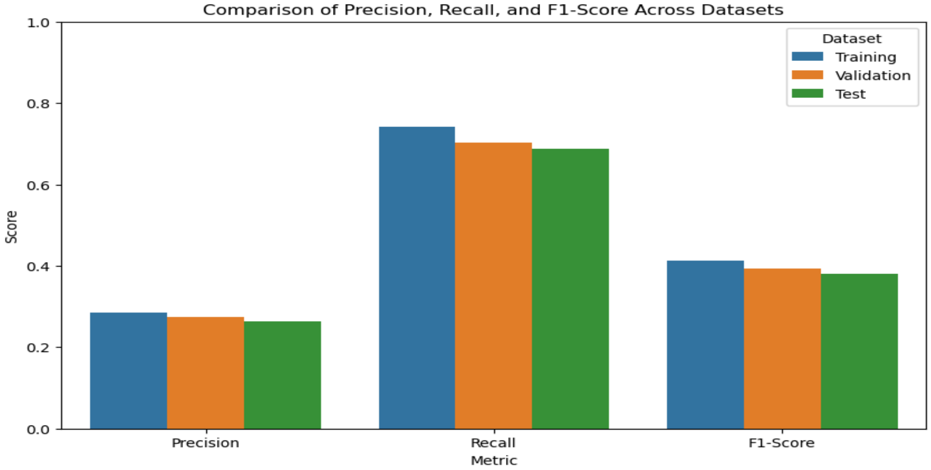


Fig- 10 (Feature Importance for Decision Trees)

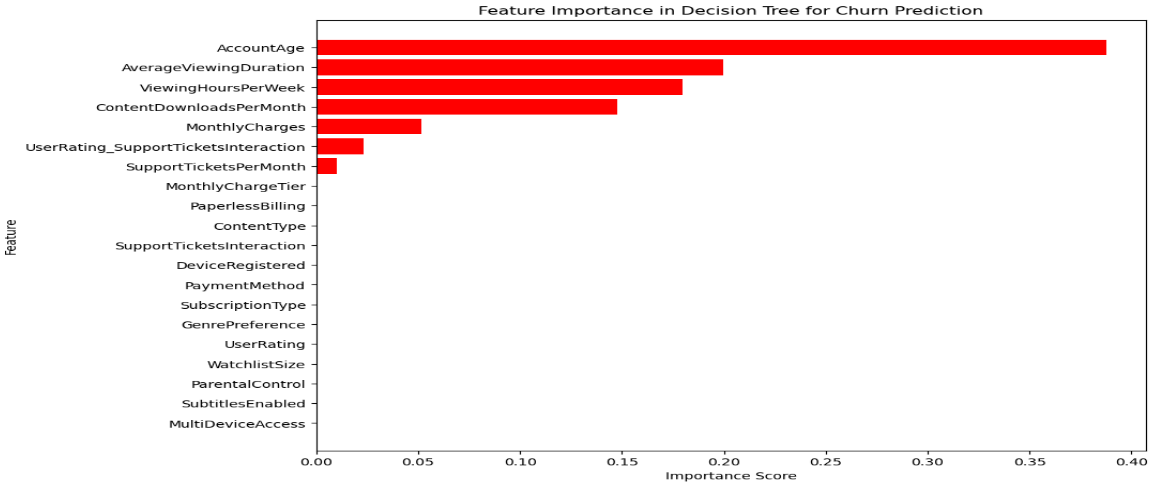


Fig- 11 (Best RandomForest Classifier model results)

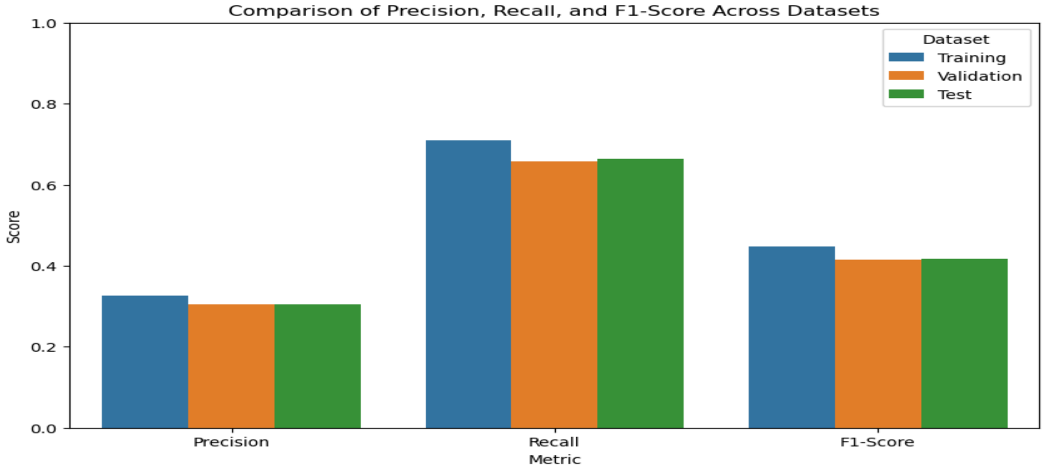


Fig- 12 (Feature Importance for RandomForest Classifier)

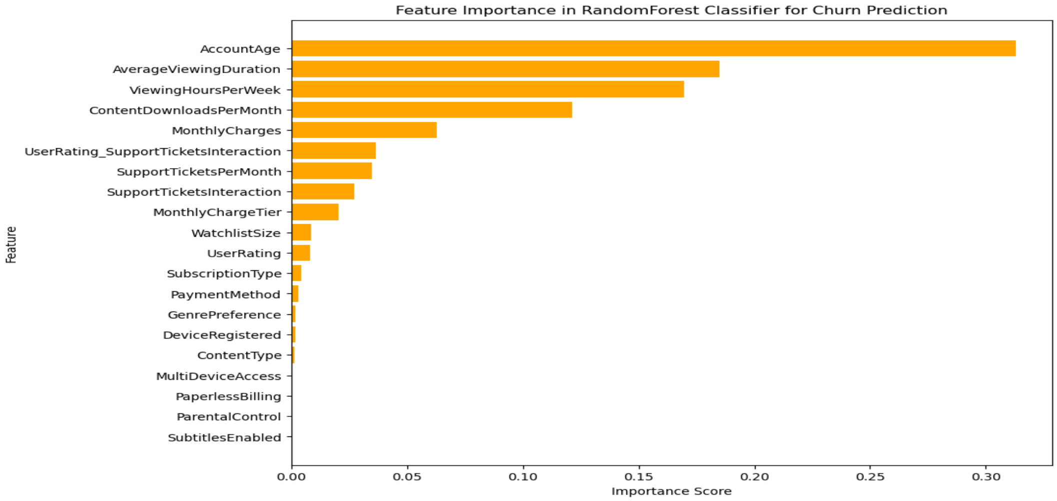


Fig- 13 (Best XGBoost Classifier model results)

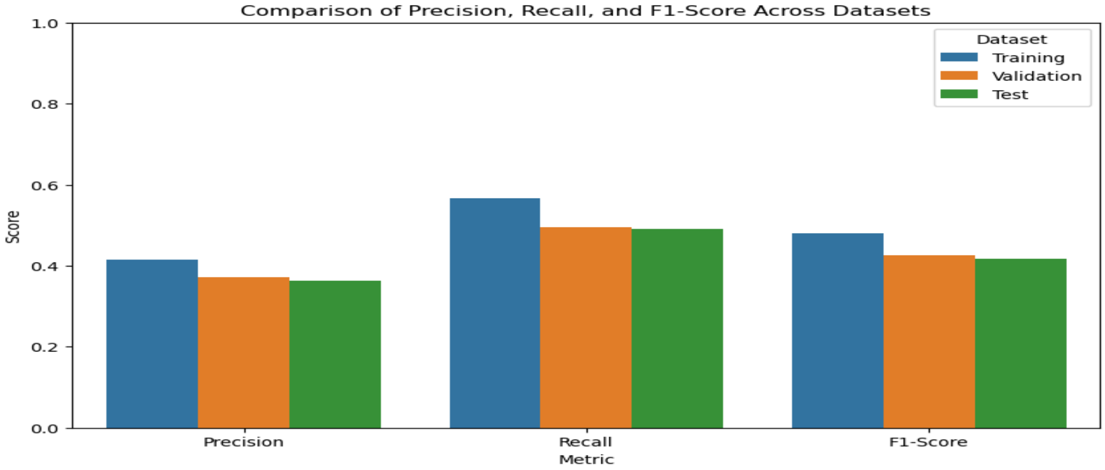
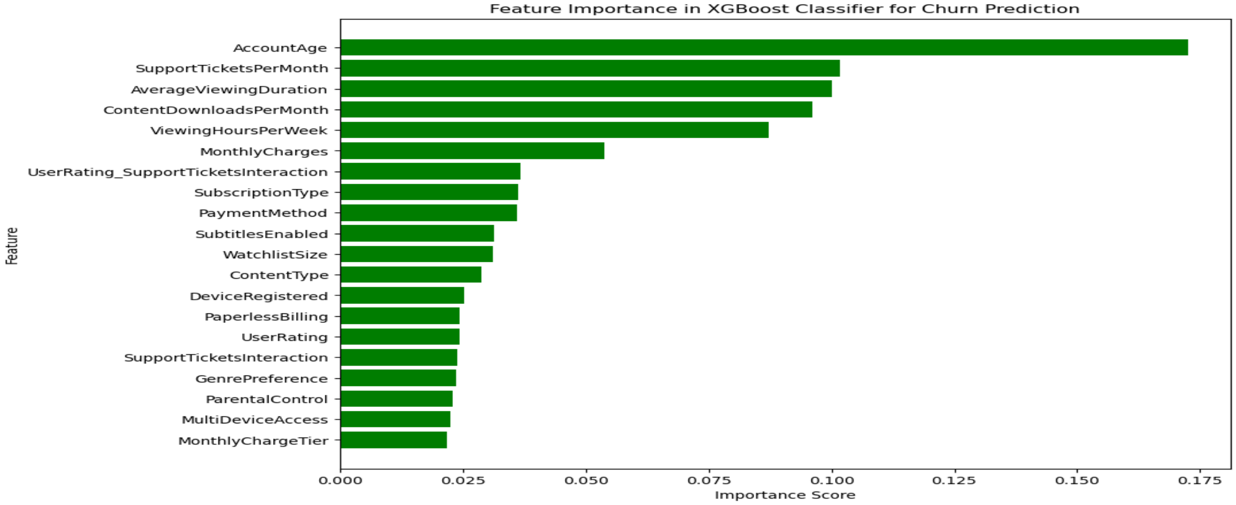


Fig- 14 (Feature Importance for XGBoost Classifier)


Feature Importance in XGBoost Classifier for Churn Prediction
AccountAge
SupportTicketsPerMonth
AverageViewingDuration
ContentDownloadsPerMonth
ViewingHoursPerWeek
MonthlyCharges
UserRating_SupportTicketsInteraction
SubscriptionType
PaymentMethod
SubtitlesEnabled
WatchlistSize
ContentType
DeviceRegistered
PaperlessBilling
UserRating
SupportTicketsInteraction
GenrePreference
ParentalControl
MultiDeviceAccess
MonthlyChargeTier
Feature
Importance Score