Predicting Customer Churn for TNB: A Machine Learning Approach to Enhance Retention Strategies

Problem Statement

TNB is a hypothetical telecommunications company used in this project to simulate a real-world scenario. The objective is to develop a predictive model that can identify customers likely to churn, enabling the company to implement proactive retention efforts.

Objectives

This project aims to develop a predictive model for TNB that will:

1. **Increase Churn Prediction Accuracy**: Through extensive communication with TNB, *I advised implementing a custom metric that aligns with their business goal*, placing higher weight on recall (0.65 * recall + 0.35 * F1-score) to prioritize identifying at-risk customers. While TNB's current performance on this metric is 0.45, the goal is to raise it to at least 0.65.
2. **Identify Key Churn Drivers**: Analyze data to uncover patterns and features most strongly linked to churn, enabling targeted strategies.
3. **Enable Proactive Retention Efforts**: Empower TNB to reach at-risk customers earlier, with more effective interventions to improve loyalty.

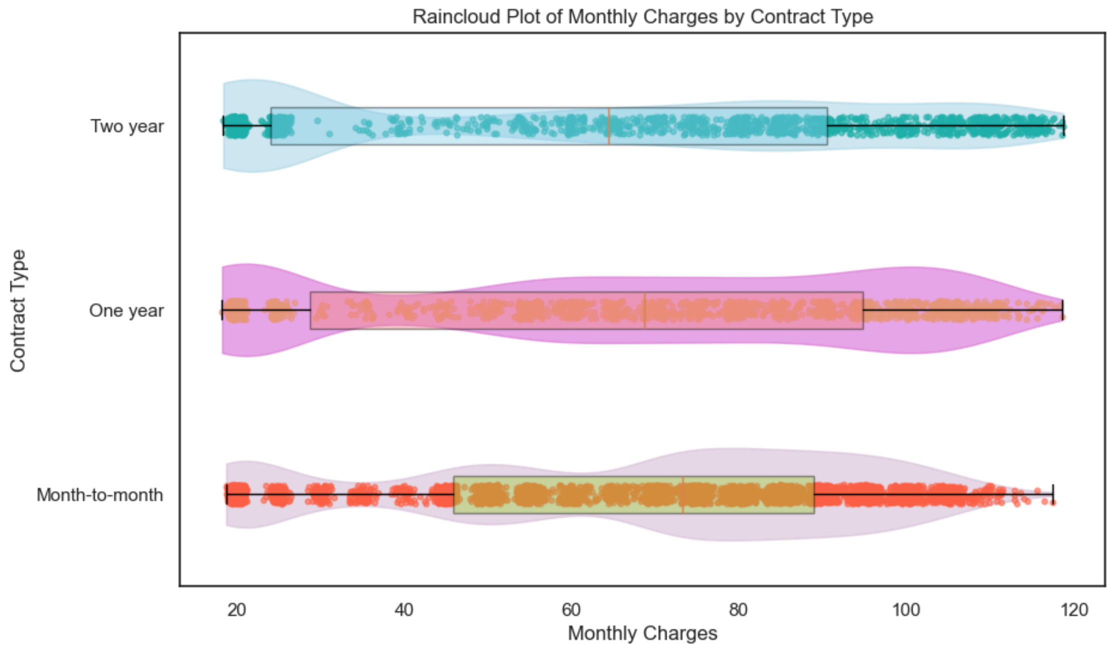
Data Overview

1. **Data Source**: This dataset is provided by TNB's Data Storage and Management Department, simulating real-world data collected on customer demographics, service usage, and account information.
2. **Dataset Structure**: The dataset includes 7,043 records across multiple columns, capturing details like demographics, contract type, payment methods, service usage, and churn status.
3. **Data Quality and Preprocessing**:
    - **Null Values**: 11 missing values were identified in the `TotalCharges` field due to a technical error during data collection, confirmed by TNB's Data Management Department. Given the low number of missing values, these rows were dropped to ensure accuracy in visualizations and analyses.
    - **Class Imbalance**: The dataset is slightly skewed, with around 5,000 non-churn and 2,000 churn cases, which reflects the typical imbalance in churn prediction tasks. Strategies to address this imbalance were applied during modeling to improve prediction reliability.
4. **Key Features**:
    - **Customer Demographics**: Age,gender, tenure.
    - **Service and Account Information**: Contract type, payment method, and monthly charges.
    - **Behavioral Data**: Service usage patterns potentially linked to churn.

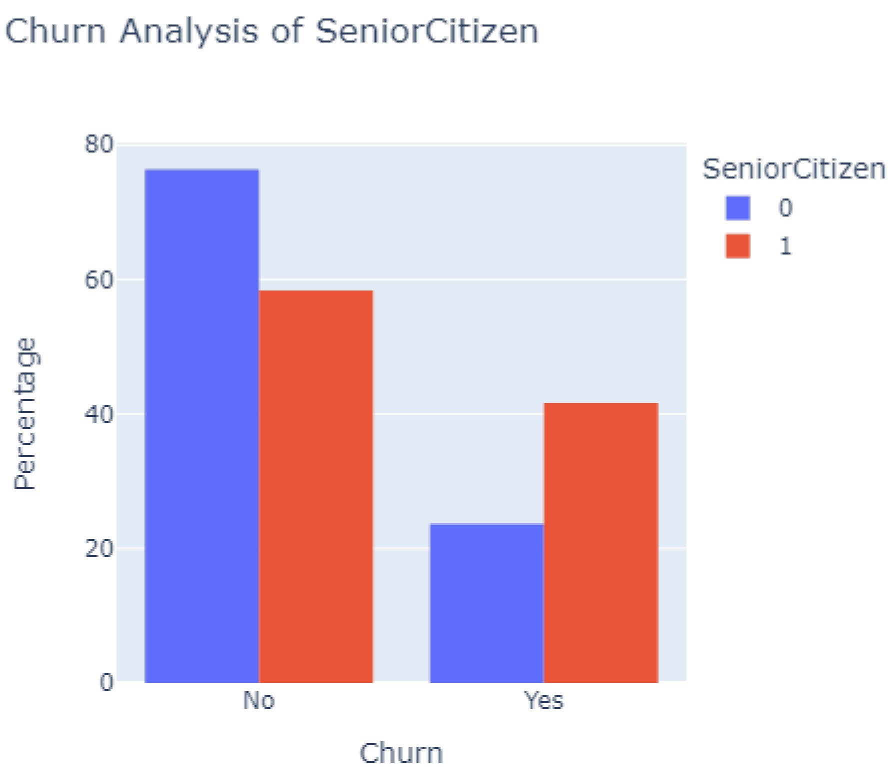
Data Preprocessing and Exploration

1. **Data Cleaning**:
    a. **Missing Values**: Various imputation techniques were explored to handle the 11 missing values in `TotalCharges`, but they proved ineffective in maintaining data integrity, so these rows were removed. This decision minimized visualization issues and retained overall data quality.
    b. **Data Type Adjustments**: Adjusted column data types, particularly converting `TotalCharges` to a numeric format, to ensure compatibility with analysis tools.

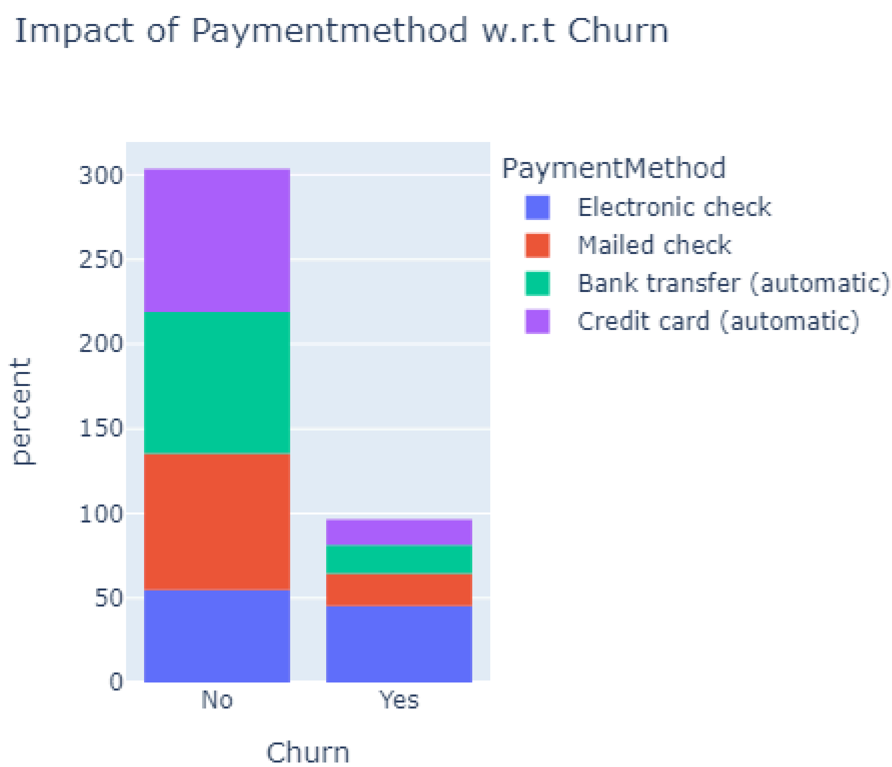
2. **Class Imbalance**: The dataset's churn class is skewed, with 5,000 non-churn and 2,000 churn records. Resampling methods like SMOTE were considered to balance the data and improve the model's predictive ability on the minority class.
3. **Distribution and Anomaly Detection**: Distributions of each attribute were thoroughly analyzed to detect any anomalies, outliers, or inconsistencies. This exploration confirmed expected patterns, such as higher charges for certain customer segments, and highlighted key areas where customer behavior varies.
4. **Key Findings from EDA**:
    a. **Charges and Churn**: Higher churn rates were observed among customers *with high monthly charges, particularly for those on month-to-month contracts.* This trend suggests that cost-sensitive customers may be more likely to leave due to monthly pricing fluctuations



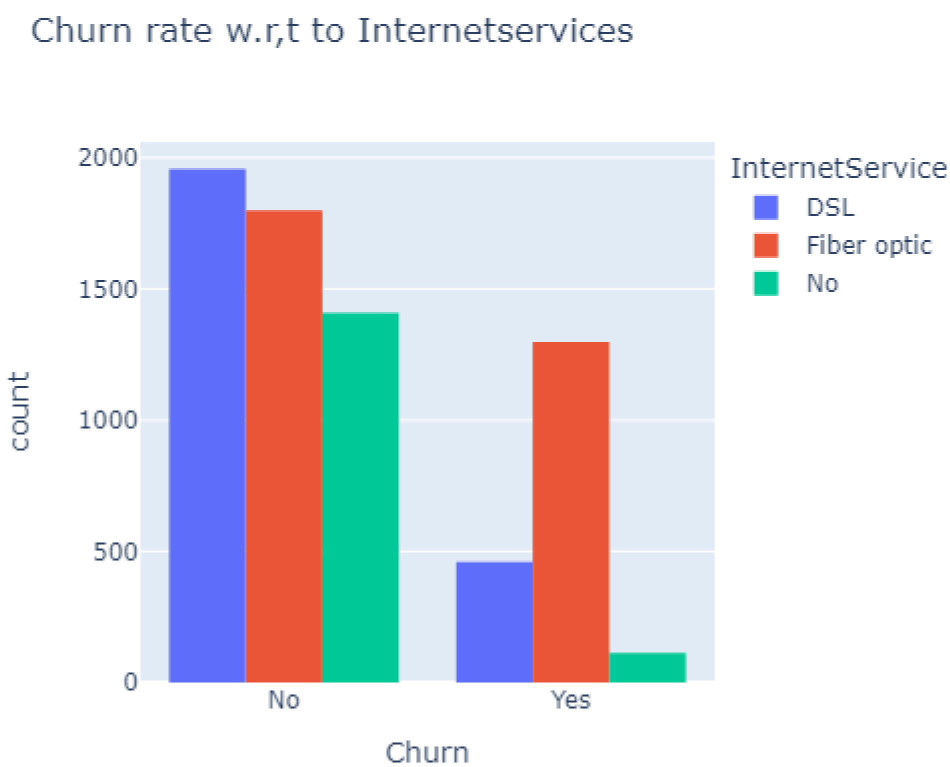
b. **Senior Citizens**: _Senior citizens exhibited a churn rate nearly double that of younger customers_. This may reflect a cautious financial approach, where non-essential services are frequently reconsidered.



c. **Automatic Payment Method**: *Customers using automatic payment methods churned less*, likely due to the reduced friction of service renewal. Convenience in billing appears to increase retention



d. **Fiber Optic Service**: *There is a high churn rate among customers using fiber optic services*, potentially due to dissatisfaction with reliability, speed, or support. This insight could guide TNB in improving service quality to reduce churn.



Modeling and Evaluation

1. **Feature Engineering**:
    - **Automated Feature Engineering**: Leveraged the FeatureTools library to automatically generate features, resulting in enhanced predictive attributes. _Most of the top 15 features were created through this process, confirming the utility of automated feature engineering for this dataset._
2. **Data Preprocessing**:
    - **Normalization and Encoding**: Applied StandardScaler to normalize continuous features and one-hot encoding for categorical features, as the dimensionality was manageable without significant computational overhead.
    - **Class Imbalance Mitigation**: Addressed class imbalance by assigning higher weights to the minority churn class across all algorithms, ensuring fair evaluation of both classes.
    - **Stratified Data Splitting**: _Used stratified splits_ to create training, validation, and test sets, maintaining class distribution and ensuring reliable model evaluation.
3. **Model Experimentation**:
    - **Initial Models**: Started with baseline models, including Logistic Regression and SVM, to gauge initial performance. These models were computationally inexpensive but yielded suboptimal results, motivating the need for more complex models.
    - **Advanced Models**: Experimented with a range of models, including **XGBoost, LightGBM, CatBoost, and Artificial Neural Networks (ANN)**. Given the low-dimensionality, numerous experiments were feasible without excessive computational costs.

    - **Hyperparameter Tuning and Ensembling**: *Used Optuna for hyperparameter tuning across multiple algorithms*, and further experimented with ensembles and optimized ensembles, though these didn't yield significant improvements.
4. **Model Selection**:
    - **Custom Metric**: To address the business need, a custom metric combining recall and F1-score was used: **Custom Metric = (0.65 * Recall) + (0.35 * F1)**. *This custom metric prioritized recall to catch more potential churn cases.*
    - **Final Model**: After extensive experimentation, the **Artificial Neural Network (ANN)** emerged as the top performer on the custom metric, achieving optimal recall and F1-score balance.

Final Model Performance

The Artificial Neural Network (ANN) model demonstrated strong performance on key metrics, indicating its suitability for predicting customer churn for TNB:

- **Custom Weighted Recall**: *0.73 — Reflects the model's performance on a recall-weighted metric aligned with TNB's goal, emphasizing recall while incorporating F1 as well.*
- **AUC**: 0.85 — Shows the model's ability to distinguish between churners and non-churners effectively.
- **Recall**: 0.80 — Ensures a high rate of correctly identified churners, addressing the business objective to capture more potential churn cases.
- **Precision**: 0.53 — Indicates the proportion of true churners among those predicted to churn, a reasonable rate given the emphasis on recall.
- **F1-Score**: 0.64 — Balances precision and recall, demonstrating the model's reliability in predicting churn.
- **Accuracy**: 0.76 — Provides a general sense of model correctness across both classes.

Post-Deployment and Business Impact

1. **Deployment Scripts and Automation**:
    - Created deployment scripts to streamline data processing, enabling the data to *be **training-ready** for continuous model retraining as new data is collected*. This ensures the model remains updated and continues to provide relevant predictions over time.
    - *Set up automated retraining schedules based on model drift detection,* allowing TNB to maintain high prediction accuracy without manual intervention. By focusing on automation, I aimed to make TNB's churn prediction process both efficient and sustainable.

2. **Cross-Department Collaboration**:
   - Beyond modeling, I prioritized communication with key departments, transforming predictive insights into actionable strategies that support TNB's business goals. Here are the identified factors contributing to churn, along with potential solutions developed in collaboration with respective departments:

   - **High Charges for Monthly Users**: Monthly users with higher costs tend to churn more. To address this, *TNB will offer* ***targeted short-term discounts*** *to customers identified as potential churners by the model*, providing a proactive approach to retain cost-sensitive users.

   - **Senior Citizen Churn**: Senior citizens exhibit a significantly higher churn rate, likely due to their conservative financial approach. *TNB's customer support team will now* ***engage with senior customers***, conducting personal outreach to understand their concerns and reduce churn through tailored support.

   - **Automatic Payment Methods**: Customers using automatic payment options have lower churn rates. *The* ***marketing team will emphasize the convenience and benefits of automatic payments***, using targeted communication strategies to increase adoption and thereby improve customer retention.

   - **Fiber Optic Service Issues**: A high churn rate among fiber optic users suggested dissatisfaction with service quality. After coordinating with the internet service team, *We identified* ***outsourced equipment as the root cause of slow internet speeds***. The company will now work with suppliers to address equipment issues and improve service quality, which could significantly reduce churn among fiber optic customers.

3. **Future Enhancements and Monitoring**:
   - The deployment setup includes tracking model performance and monitoring key metrics to ensure alignment with TNB's churn reduction goals.

   - Plans for continuous improvement involve expanding the predictive model to offer additional customer insights, such as likely engagement with retention offers or sensitivity to price changes, to further enhance retention strategies.