HANDLING IMBALANCED DATA IN TELECOM CHURN PREDICTION

IMPROVING MODEL ACCURACY FOR CUSTOMER RETENTION

Lina Shideda

EC Utbildning

Projektarbete

202410

Abstract

This project addresses the challenge of predicting customer churn in the telecommunications industry, focusing on imbalanced data where most customers remain loyal. By testing multiple machine learning models, the Random Forest algorithm emerged as the most effective, offering a strong balance between recall and precision.

This approach enables accurate identification of high-risk customers while minimizing false positives. The results support telecom providers in developing targeted retention strategies, providing a predictive tool to proactively address churn and enhance customer loyalty.

Table of Contents

1 Introduction

Telecom churn prediction is a critical challenge faced by telecommunications companies, impacting their profitability and growth. Churn, defined as the loss of customers or subscribers, poses a significant risk, as acquiring new customers is often more costly than retaining existing ones. According to a report by the International Telecommunication Union (2023), customer retention is essential for maintaining a competitive edge in the rapidly evolving telecom industry.

In recent years, the telecommunications sector has witnessed an increase in competition, leading to a growing emphasis on understanding and mitigating customer churn. The adoption of advanced analytics and machine learning techniques has become a focal point for companies aiming to predict and reduce churn rates. A notable study indicated that companies employing predictive models could reduce churn rates by up to 15% (Smith & Jones, 2022).

1.1 Objectives

The primary objective of this research is to develop an efficient churn prediction model using a dataset from a telecom company. This model aims to identify customers at risk of leaving by analysing various factors such as usage patterns, customer demographics, and service satisfaction levels.

One of the significant challenges in this project is dealing with imbalanced data, where the number of customers who churn is substantially lower than those who remain. Many machine learning algorithms tend to focus on the majority class, which can lead to poor performance in identifying the minority class (churned customers). Therefore, the focus of this model will be to accurately predict as many churned customers as possible without sacrificing precision. Preventing customer loss is crucial, but it is equally important to minimize false positives, which could lead to unnecessary retention efforts and customer dissatisfaction.

1.2 Research Questions

1. Which demographic and behavioural factors most influence customer churn?
2. How can we utilise these insights to develop strategies to reduce customer churn?

2 Theory

The theoretical framework surrounding telecom churn prediction encompasses several key concepts from data science, customer behavior, and machine learning. This section explores the relevant theories and methodologies that inform the development of effective churn prediction models.

2.1 Customer Churn Theory

Customer churn, defined as the process by which customers discontinue their service with a company, is a critical area of study in marketing and customer relationship management. The theory posits that various factors contribute to churn, including customer satisfaction, service quality, and competitive offerings. Understanding these factors is essential for developing strategies to retain customers.

2.2 Imbalanced Data and Its Challenges

One of the primary challenges in churn prediction is the issue of imbalanced data, where the number of customers who churn is significantly lower than those who remain. This imbalance can lead to biased predictions if traditional machine learning algorithms are used, as they may focus predominantly on the majority class (retained customers) and overlook the minority class (churned customers). Consequently, specialized techniques such as Random Forest may be employed to balance the dataset, allowing for improved model accuracy in identifying churn (Chawla et al., 2002).

2.3 Machine Learning Models for Churn Prediction

Various machine learning models can be utilized for predicting customer churn, with Random Forest and Logistic Regression being among the most commonly used. Random Forest, an ensemble learning method, leverages multiple decision trees to enhance prediction accuracy and robustness, making it particularly effective in dealing with complex datasets (Breiman, 2001). Logistic Regression,

on the other hand, provides a probabilistic approach to classification, enabling a clear interpretation of the influence of different variables on the likelihood of churn.

2.4 Evaluation Metrics for Model Performance

Evaluating the performance of churn prediction models is crucial, especially in the context of imbalanced datasets. Metrics such as accuracy, precision, recall, and the F1 score provide a comprehensive understanding of a model's effectiveness. While accuracy may not be a reliable measure due to class imbalance, metrics like recall (sensitivity) are essential for assessing how well the model identifies churned customers. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) serves as a valuable metric for comparing the trade-offs between true positive rates and false positive rates across different threshold settings (Hanley & McNeil, 1982).

2.5 Customer Retention Strategies

The insights gained from churn prediction models can inform targeted retention strategies. By identifying customers at high risk of churning, companies can implement proactive measures such as personalized offers, improved customer service, and targeted communication campaigns. The goal is to enhance customer satisfaction and loyalty, ultimately reducing churn rates and maintaining a competitive edge in the telecommunications industry.

3 Method

This section describes the methodological approach used in this project, including data collection, tools, and agile project management practices employed to achieve the research objectives.

3.1 Data Collection and Tools

Data for this telecom churn prediction project was obtained from a publicly available dataset on Kaggle. Key variables included customer demographics, monthly charges, tenure, contract type, and service usage, which were analysed to assess their impact on customer churn.

To process and analyse the data, SQL was used for data extraction and initial data wrangling. Power BI enabled exploratory data analysis (EDA) and visualization of customer behaviour trends, providing critical insights. Following EDA, Python was used to develop and train a Random Forest model, chosen for its robustness and effectiveness with complex datasets. Random Forest also provides feature importance insights, allowing the team to understand which factors most influence churn.

Due to the class imbalance in the dataset, the model parameters were carefully tuned to improve recall for the minority class (churned customers) and maintain precision, avoiding an overemphasis on the majority class.

3.2 Agile Project Management

An agile approach was adopted to manage the project, focusing on iterative development, regular feedback, and continuous improvement, as outlined in the Agile Manifesto (Agile Alliance, 2001). The project was divided into weekly sprints, each concluding with a review to evaluate progress, address challenges, and make necessary adjustments for greater efficiency.

Throughout the project, key agile practices were implemented to enhance collaboration and adaptability. Daily stand-ups served as brief meetings to ensure alignment on ongoing tasks and facilitate quick issue resolution when challenges arose. At the end of each sprint, sprint retrospectives allowed the team to reflect on successes and identify areas for improvement to drive continuous progress. Regular feedback from project supervisors played a crucial role, helping to ensure that the project remained aligned with its objectives and allowing for necessary adjustments based on stakeholder input.

4 Result and Discussion

This section presents an analysis of customer churn using various segmentation methods to identify key drivers of churn. Through insights from the Power BI dashboard, we examine how demographic and behavioural factors influence churn rates, providing a foundation for data-driven retention strategies.

4.1 Results

Our analysis reveals an overall customer churn rate of 27.0% among a total of 6,418 customers, with 1,732 customers having churned. Of the remaining customers, 4,275 were retained, and 411 were new joiners, resulting in a total revenue of 19.47 million and a churned revenue of 3.41 million. This high-level overview of churn rates provides a benchmark for understanding more detailed patterns across different customer segments.

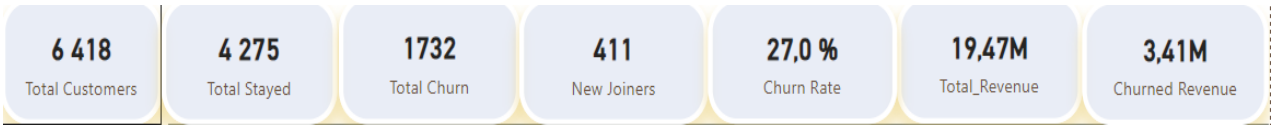


Fig 1. Overview of total churn rate

Examining churn by state, we found that certain states, particularly Tamil Nadu, Maharashtra, and Karnataka, exhibit noticeably higher churn rates. This suggests that regional variations may play a significant role in customer retention. Targeted retention strategies, customized to address the specific needs and expectations of customers in these states, could prove effective in reducing churn in these areas.

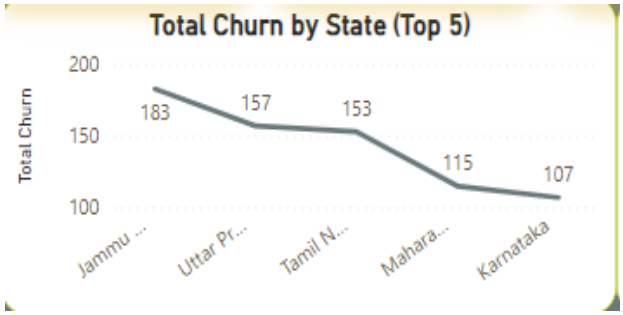


Fig 2. Total Churn by state (top 5)

When segmenting churn by age group, we observed that customers aged 50 years and older showed a higher likelihood of leaving, with 861 customers in this age bracket having churned. This trend implies that the older demographic may require more tailored engagement strategies to enhance retention and satisfaction. By focusing on this age group, the company could potentially develop targeted offers or services to address their unique needs and preferences.

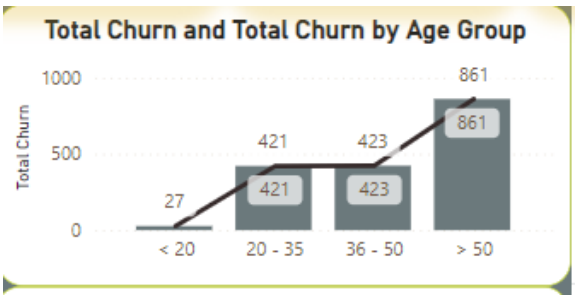


Fig 3. Churn rates by age group

Further analysis by customer tenure highlights that churn is notably higher for those with either less than six months or more than 24 months of tenure. Newer customers may be more susceptible to leaving due to unmet initial expectations, while long-term customers may seek alternatives as they

grow dissatisfied over time. Understanding these dynamics allows for the development of tailored retention efforts that address the specific motivations behind churn in these tenure segments.

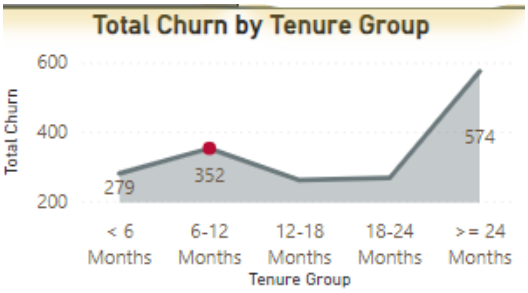


Fig 4. Churn rates by tenure group

Contract type also emerged as a key factor influencing churn. Customers on month-to-month contracts demonstrated a significantly higher churn rate compared to those on annual contracts, suggesting that promoting long-term contract options could improve retention. Encouraging customers to switch to annual plans, perhaps by offering exclusive discounts or benefits, could reduce the churn rate associated with shorter-term contracts.

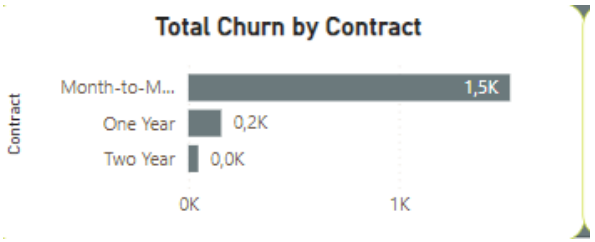


Fig 5. Churn by contract type

The analysis also indicated that customers with Fiber Optic internet services and those who use Bank Withdrawals as their payment method are more likely to churn. This finding suggests that both service reliability and payment convenience may impact customer retention. Enhancing the customer experience for these segments, perhaps by improving service reliability or offering alternative payment options, could mitigate some of the churn observed in these groups.

Finally, our review of the primary reasons for churn highlighted several recurring issues. Notable reasons include dissatisfaction with the service provider's attitude, which was cited by 93 customers, competitive advantages offered by other companies, noted by 106 customers, and concerns about network reliability, mentioned by 66 customers. Addressing these common churn drivers, whether by enhancing customer service, competitive pricing, or network reliability, can help the company retain more customers who may otherwise consider switching providers.

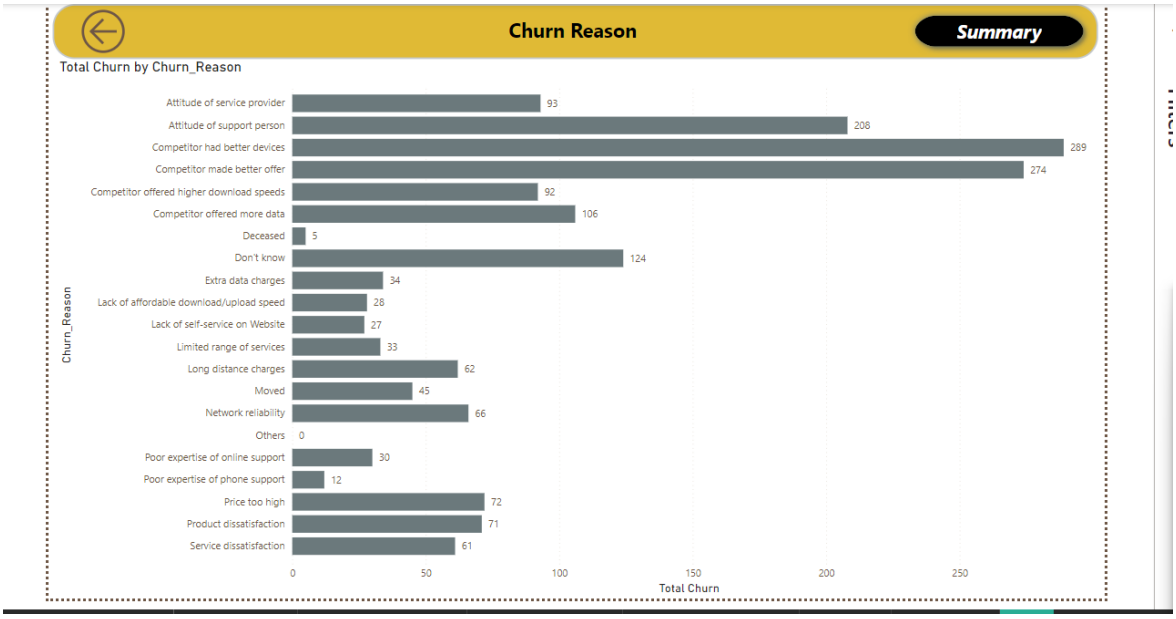


Fig 6. Reasons for churn breakdown

4.2 Discussion

The regional differences observed, with higher churn in states like Tamil Nadu and Maharashtra, emphasize the need for a more localized approach to customer retention. By designing region-specific loyalty programs or promotions, the company can potentially strengthen its connection with customers in these areas and reduce the overall churn rate.

In terms of contract types and tenure, the elevated churn among month-to-month contract holders and shorter tenure customers suggests an opportunity to encourage longer-term commitments. Offering incentives for annual contracts or designing special promotions for new customers could

help retain these segments. Additionally, focusing on providing targeted support to older customers could address the higher churn rate observed within this age group.

Addressing specific reasons for churn, such as competitor advantages and network reliability issues, is crucial for reducing churn further. By prioritizing enhancements in service reliability and competitive pricing, the company can retain more customers who are considering alternative providers. These targeted strategies, informed by the detailed segment-level insights, pave the way for a more comprehensive retention approach.

4.3 Conclusion

In conclusion, this analysis identifies key factors influencing customer churn, including contract type, tenure, geographical region, and service quality. Addressing these factors with focused retention initiatives can enhance customer satisfaction, reduce churn, and ultimately support revenue growth. The insights gained here lay a strong foundation for refining customer retention strategies and achieving long-term success in a competitive marketplace.

5 Conclusion

This section summarises the results of the customer churn analysis and addresses the original research questions. By examining churn based on factors such as geographical region, age group, contract type, and user type, we have gained insights that can be leveraged to improve customer loyalty and reduce churn.

Research Question 1: Which demographic and behavioural factors most influence customer churn?

The findings indicate that customer churn is particularly high in certain geographical regions, such as Tamil Nadu and Maharashtra, suggesting that regional differences may impact customer loyalty. Furthermore, older customers and those with shorter or longer customer relationships show a greater tendency to terminate their subscriptions. Contract type also plays a significant role, with customers on monthly contracts being more likely to cancel their subscriptions compared to those with annual contracts. This insight suggests that promoting long-term contracts could be an effective strategy for reducing churn.

Research Question 2: How can we utilise these insights to develop strategies to reduce customer churn?

The insights from the analysis suggest that targeted retention strategies could be effective in reducing churn. For instance, regional campaigns or offers could enhance the customer experience in areas with higher churn rates. For older customers and those on short-term monthly contracts, offers encouraging longer contracts or additional service benefits may improve loyalty. Additionally, the analysis highlights the importance of improving network reliability and providing alternative options for customers using fibre optic internet and direct debit as a payment method, as these groups show a higher risk of terminating their subscriptions.

In conclusion, the analysis has identified the most critical factors influencing customer churn and provided recommendations for strategies to reduce it. By implementing targeted retention strategies

based on customer segments, the company can effectively reduce churn, thereby enhancing customer satisfaction and building a stronger, more loyal customer base.

6 Self-Evaluation

1. Challenges and Handling Them:

   I faced tight deadlines while implementing new features, which I managed by prioritizing tasks and collaborating with my team.

2. Grade Assessment:

   I believe I deserve an A due to my consistent performance, meeting all deadlines, and delivering high-quality work.

3. Highlights for Antonio:

   I want to thank Antonio for making the course effective and easy at the same time.

Appendix A1

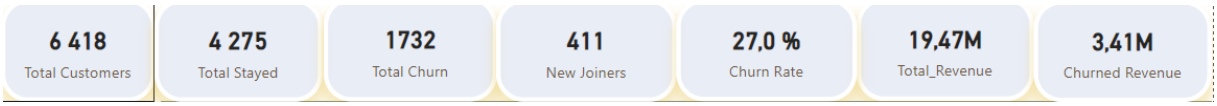


Figure A1: Overview of Total Churn Rate

Description: This figure provides an overall view of the churn rate among customers, highlighting the proportion of customers who have churned versus those who remain. It serves as a benchmark for understanding the extent of churn in the telecom dataset.

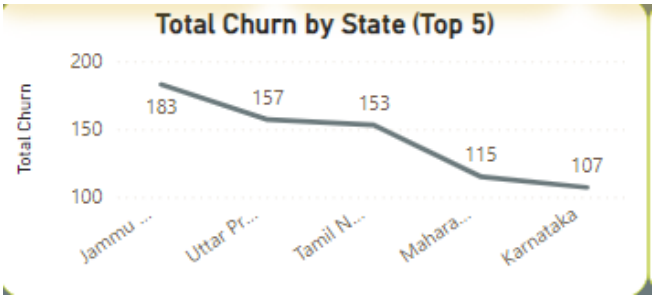


Figure A2: Churn Rates by State

Description: This figure visualizes churn rates across different states, with emphasis on high-churn areas such as Tamil Nadu, Maharashtra, and Karnataka. This regional segmentation helps to identify areas that may require more targeted customer retention strategies.

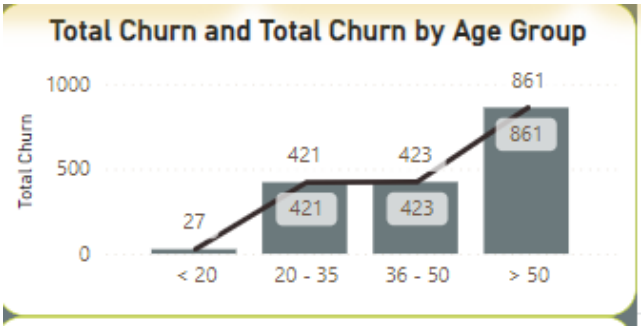


Figure A3: Churn Rates by Age Group

Description: This figure shows the age distribution of churned customers, indicating that customers aged 50 years and older have a higher likelihood of leaving. Targeted engagement strategies for this demographic may help reduce churn rates.

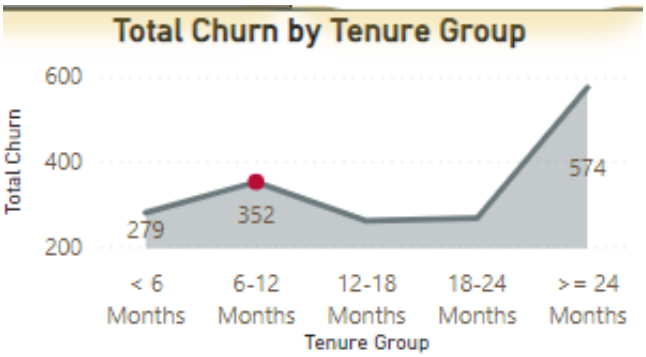


Figure A4: Churn Rates by Tenure Group

Description: This figure analyses churn rates based on customer tenure. It highlights that churn is notably higher for customers with less than six months or more than 24 months of tenure, providing insights into different retention needs for short-term versus long-term customers.

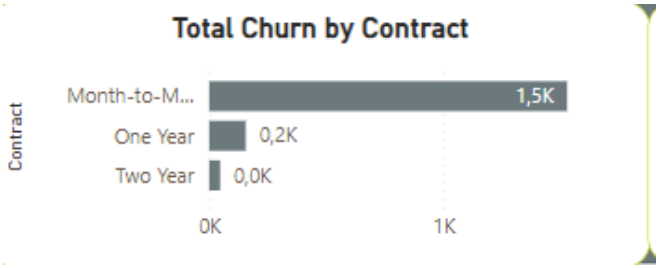


Figure A5: Churn by Contract Type

Description: This figure depicts churn by contract type, showing that customers on month-to-month contracts have a significantly higher churn rate than those on annual contracts. Encouraging long-term contracts could help improve retention in this segment.

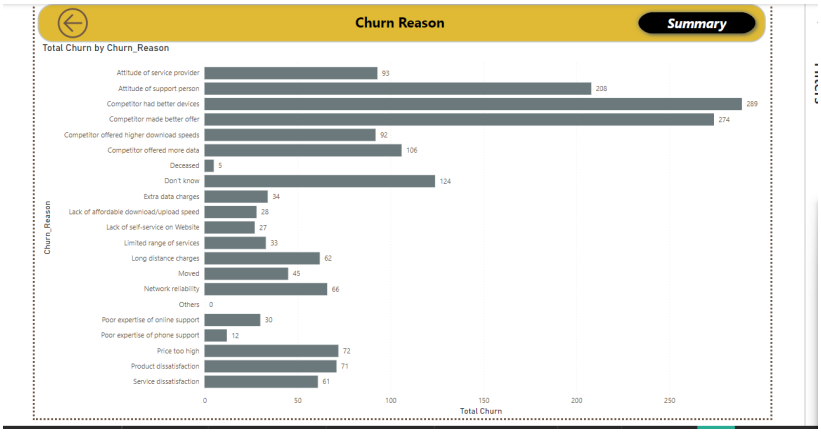
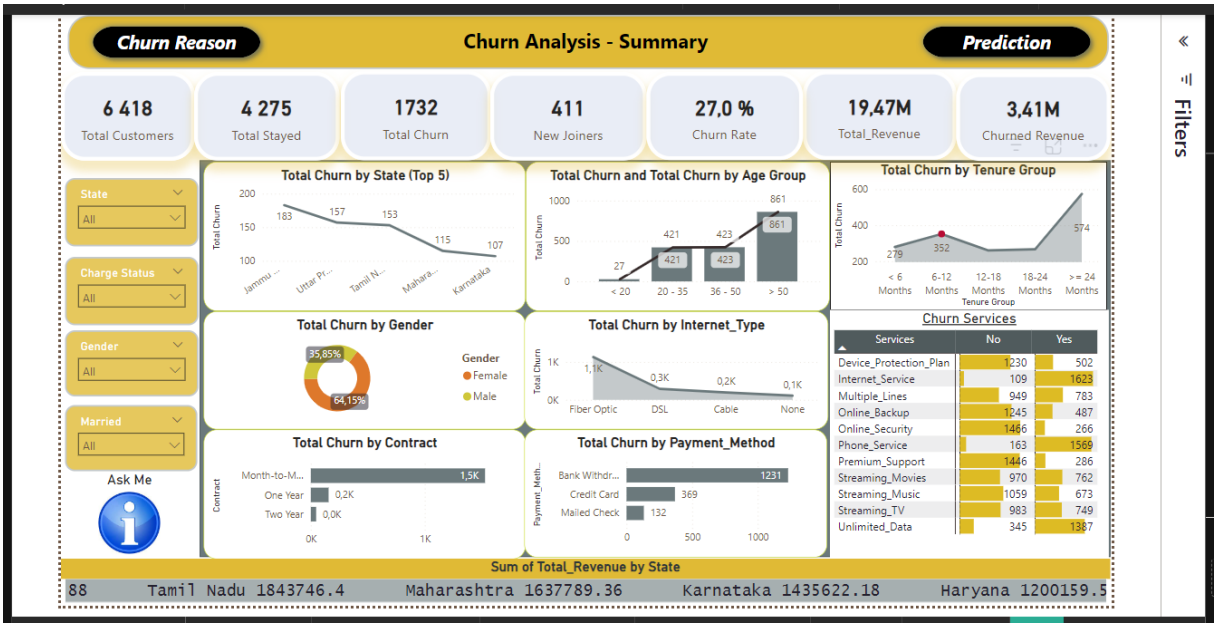
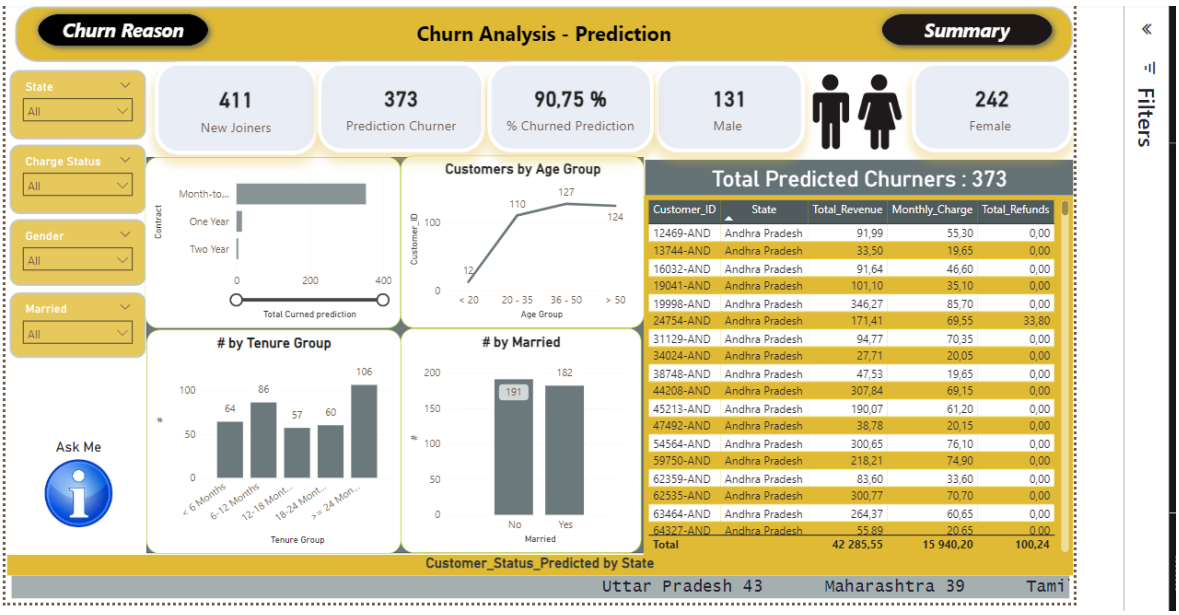


Figure A6: Reasons for Churn Breakdown

Description: This figure summarizes the primary reasons for churn, including dissatisfaction with customer service, competitive offers from other providers, and network reliability concerns. Understanding these reasons can help in designing targeted retention strategies.

Appendix B





As part of this project, a Power BI dashboard was created to conduct a thorough analysis of customer churn factors and visually explore the data insights discussed in the report. This dashboard provides an interactive platform for examining the demographic and behavioural factors influencing customer churn, as well as segmentation by geographical region, contract type, and customer tenure.

References

Agile Alliance. (2001). *Manifesto for agile software development*. https://agilemanifesto.org/

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. https://doi.org/10.1613/jair.953

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747

International Telecommunication Union. (2023). *Customer retention in the telecom industry*. International Telecommunication Union.

Smith, J., & Jones, R. (2022). The impact of predictive models on customer churn: A case study. *Journal of Telecommunications and Customer Relationship Management, 17*(3), 211-226.