BBA Semester – VI
Research Project

Name	Harsh Kumar Pandit
USN	212VBBR00456
Elective	DATA SCIENCE AND ANALYTICS
Date of Submission	02/09/2024

A study on “Customer Churn”

Research Project submitted to Jain Online (Deemed-to-be University)
In partial fulfillment of the requirements for the award of:

Bachelor of Business Administration

Submitted by:

Harsh Kumar Pandit

USN:

212VBBR00456

Under the guidance of:

Milind Desai

(Faculty-JAIN Online)

Jain Online (Deemed-to-be University)
Bangalore

2023-24

DECLARATION

I, *Harsh Kumar Pandit* hereby declare that the Research Project Report titled *"(Customer Churn)" has been* prepared by me under the guidance of the *Milind Desai.* I declare that this Project work is towards the partial fulfillment of the University Regulations for the award of the degree of Bachelor of Business Administration by Jain University, Bengaluru. I have undergone a project for a period of Six Weeks. I further declare that this Project is based on the original study undertaken by me and has not been submitted for the award of any degree/diploma from any other University / Institution.

Place: Harsh Kumar Pandit

Date: *Name of the Student*

USN: 212VBBR00456

CERTIFICATE

This is to certify that the Research Project report submitted by Mr./Ms. *Harsh Kumar Pandit* bearing *(*212VBBR00456*)* on the title *"Customer Churn"* is a record of project work done by him/ her during the academic year 2022-23 under my guidance and supervision in partial fulfillment of Bachelor of Business Administration.

Place:

Date: *Faculty Guide*

ACKNOWLEDGEMENT

The Learners may acknowledge the organization guide, University officials, faculty guide, other faculty members, and anyone else they wish to thank for their contribution towards accomplishing the research project successfully. The Learners may write in their own words and in small paragraph.

Harsh Kumar Pandit

Name of the Student

*USN:*212VBBR00456

EXECUTIVE SUMMARY

In today's competitive E-commerce and DTH markets, customer retention is more critical than ever. This project addresses the pressing challenge of predicting customer churn—a phenomenon where customers discontinue their services, often leading to significant revenue loss. The primary objective is to develop a predictive model that identifies accounts at high risk of churn, enabling the company to implement targeted retention strategies that are both effective and efficient.

The project begins by defining the business problem, emphasizing the importance of customer retention in safeguarding revenue and optimizing resources. Given the nature of these markets, where an account often represents multiple users, the loss of even a single account can have a substantial financial impact. Therefore, the ability to predict churn and proactively address it provides the company with a significant business advantage.

To tackle this challenge, we collected and analyzed a dataset containing various customer attributes, including demographic information, service usage, customer care interactions, and revenue-related metrics. The data spans a substantial period, ensuring that the model is trained on a comprehensive set of customer behaviors and interactions.

The project employs a rigorous Exploratory Data Analysis (EDA) to understand the underlying patterns and relationships within the data. This includes analyzing continuous and categorical variables, identifying potential outliers, and treating missing values to ensure data integrity. The EDA also uncovers critical business insights, such as the impact of service scores and customer complaints on churn likelihood, which can guide targeted retention efforts.

With the data prepared, several predictive models were developed, with Logistic Regression being the primary focus. The model's performance was evaluated using various metrics, including accuracy, precision, recall, F1 score, and AUC-ROC, ensuring a robust and reliable prediction of churn. The analysis revealed that specific

factors, such as lower service satisfaction and higher complaint frequency, significantly increase the probability of churn.

The business implications of these findings are profound. By identifying high-risk accounts, the company can tailor its retention strategies to address specific customer pain points. For example, accounts with low service scores could be targeted with enhanced customer support, while those with frequent complaints might benefit from personalized outreach and problem resolution. Additionally, the insights gained from the EDA enable the company to refine its loyalty programs, ensuring that resources are allocated efficiently to maximize customer lifetime value.

In conclusion, this project provides a comprehensive approach to predicting and mitigating customer churn. The model and insights generated offer actionable recommendations that can significantly improve customer retention, reduce churn rates, and ultimately protect and enhance the company's revenue streams. By focusing on targeted, data-driven retention strategies, the company can maintain a competitive edge in the ever-evolving E-commerce and DTH markets.

TABLE OF CONTENTS

List of Tables

List of Graphs

CHAPTER 1

INTRODUCTION AND BACKGROUND

INTRODUCTION AND BACKGROUND

1.1 Purpose of the Study

The purpose of this study is to develop a churn prediction model for an e-commerce or DTH (Direct-to-Home) provider. By analyzing customer data, including variables such as tenure, service scores, and revenue metrics, the goal is to identify potential churners—customers who are likely to discontinue their service. This model aims to segment these potential churners and provide targeted campaign recommendations to enhance customer retention while balancing profitability. The insights gained from this study will help in designing strategies that reduce churn without resorting to excessive discounts or free offers, thereby optimizing revenue.

1.2 Introduction to the Topic

Customer churn is a critical issue in the e-commerce and DTH industries, where retaining existing customers is often more cost-effective than acquiring new ones. Churn prediction models use historical data to forecast which customers are at risk of leaving. Understanding the factors that contribute to churn enables companies to implement proactive measures to retain these customers. This study employs data analysis techniques and machine learning models to predict churn and offers actionable insights for improving customer retention strategies.

1.3 Overview of Theoretical Concepts

- Churn Prediction: This involves using statistical and machine learning methods to identify customers who are likely to churn based on their behavior and attributes.
- Feature Engineering: The process of selecting and transforming variables to improve model performance. Key features in churn prediction might include customer tenure, service usage scores, and revenue metrics.
- Model Evaluation: Assessing the performance of the churn prediction model using metrics such as accuracy, precision, recall, and F1 score to ensure the model's effectiveness in predicting churn.

1.4 Company/ Domain / Vertical /Industry Overview

- E-commerce Industry: A sector focused on buying and selling goods and services online. Key players in this industry leverage data analytics to understand consumer behavior, predict churn, and enhance customer engagement.

- DTH Industry: Involves providing television programming via satellite signals. Companies in this industry face challenges such as customer retention and competitive pricing, making churn prediction crucial for maintaining a stable customer base.

1.5 Environmental Analysis (PESTEL Analysis)

- Political: Regulatory changes and government policies affecting e-commerce and media industries, including data protection laws and digital transaction regulations.
- Economic: Economic conditions influence consumer spending and subscription services. Factors like economic downturns can have affect over churn rates.
- Social: Changing consumer preferences and behaviors, including the growing demand for personalized services and digital content consumption.
- Technological: Advancements in technology such as data analytics, machine learning, and AI that enable better prediction and management of customer churn.
- Environmental: Considerations related to sustainability and corporate responsibility in business practices, which might influence customer perceptions and loyalty.
- Legal: Compliance with legal requirements related to customer data handling, privacy, and digital marketing practices.

CHAPTER 2

REVIEW OF LITERATURE

REVIEW OF LITERATURE

2.1 Domain/ Topic Specific Review

Churn Prediction Models:

- Traditional Statistical Methods: Early approaches to churn prediction often involved logistic regression models due to their interpretability and effectiveness in binary classification tasks. Studies have shown that logistic regression can effectively predict churn by analyzing customer behavior and service usage patterns.
- Machine Learning Approaches: Recent advancements in machine learning have led to the development of more sophisticated churn prediction models. Techniques such as decision trees, random forests, gradient boosting machines, and neural networks have been employed to enhance predictive accuracy. Research indicates that ensemble methods and deep learning models can capture complex patterns in customer data better than traditional methods.
- Feature Engineering and Selection: Effective feature engineering is crucial for improving model performance. Key features identified in literature include customer tenure, interaction history, service usage patterns, and customer satisfaction scores. Studies emphasize the importance of selecting relevant features to avoid overfitting and improve model generalizability.
- Evaluation Metrics: Common evaluation metrics for churn prediction models include accuracy, precision, recall, F1 score, and AUC-ROC. Research highlights the importance of using multiple metrics to assess model performance comprehensively, especially in imbalanced datasets where churners are a minority.

Customer Retention Strategies:

- Targeted Marketing: Literature suggests that personalized marketing campaigns based on churn predictions can significantly improve customer retention rates. By tailoring offers and incentives to at-risk customers, companies can reduce churn and increase customer lifetime value.
- Customer Engagement: Studies indicate that enhancing customer engagement through improved service quality, loyalty programs, and responsive customer support can mitigate churn. Effective engagement strategies are often linked to higher customer satisfaction and retention.

E-commerce and DTH Industry Insights:

- E-commerce: Research in the e-commerce domain focuses on the impact of customer behavior analytics on retention strategies. Factors such as purchase history, browsing patterns, and customer feedback are critical in predicting churn.

- DTH: In the DTH industry, studies emphasize the role of service quality and customer experience in predicting churn. Factors such as signal quality, content variety, and customer support are significant predictors of customer retention.

2.2 Gap Analysis

Existing Research:

- Model Performance: While numerous studies have explored churn prediction using various models, there is still a gap in understanding how different models perform in diverse industry contexts. Existing research often focuses on specific industries, and comparative studies across different sectors are limited.
- Feature Relevance: Although feature engineering is a well-explored area, there is a need for more research on the relevance of emerging features in churn prediction. For instance, the impact of new customer interaction channels and social media activity on churn has not been extensively studied.
- Integration of Advanced Techniques: While advanced machine learning techniques have shown promise, there is limited research on integrating these methods with traditional approaches for improved performance. Combining models and techniques to leverage their strengths remains an underexplored area.

Research Gaps:

- Industry-Specific Models: There is a lack of comprehensive studies focusing on churn prediction models tailored specifically to the e-commerce or DTH industries. Research often generalizes findings across industries, which may not capture unique industry-specific factors.
- Customer Segmentation: More research is needed on how different customer segments respond to churn prediction models and retention strategies. Understanding segment-specific behaviors and tailoring strategies accordingly could improve model effectiveness.
- Implementation and Practical Applications: There is a gap in translating theoretical models into practical, actionable strategies. Research often focuses on model development without sufficient emphasis on practical implementation and the real-world impact of these models.

CHAPTER 3

RESEARCH METHODOLOGY

RESEARCH METHODOLOGY

3.1 Objectives of the Study

- Develop a Predictive Model: To create a churn prediction model using machine learning techniques to accurately forecast customer churn in the e-commerce or DTH sector.
- Segment Potential Churners: To identify and segment customers at risk of churn based on their behavior and attributes.
- Optimize Retention Strategies: To provide actionable recommendations for retention campaigns that balance customer retention with profitability.
- Evaluate Model Performance: To assess the effectiveness of different machine learning algorithms in predicting churn and select the most suitable model.

3.2 Scope of the Study

- Industry Focus: The study focuses on the e-commerce or DTH industry, analyzing customer churn data specific to this sector.
- Data Coverage: The analysis covers customer attributes, service usage patterns, and other relevant features as outlined in the dataset.
- Geographical Scope: The study may be limited to a specific region or demographic, depending on the dataset used.
- Time Frame: The study uses historical data to predict future churn, with a focus on recent trends and patterns.

3.3 Methodology

3.3.1 Research Design

- Type: The research employs a quantitative approach, utilizing statistical and machine learning methods to analyze churn data.
- Modeling: The study involves developing and comparing various predictive models to determine the most effective approach for churn prediction.
- Evaluation: Performance metrics such as accuracy, precision, recall, F1 score, and AUC-ROC are used to evaluate model effectiveness.

3.3.2 Data Collection

- Source: Data is collected from the dataset located at C:\Extra\Capstone Project\Fresh\Customer+Churn+Data.csv.
- Variables: The dataset includes features such as 'Churn', 'Tenure', 'City_Tier', 'CC_Contacted_LY', 'Service_Score', 'Account_user_count', 'CC_Agent_Score', 'rev_per_month', 'Complain_ly', 'rev_growth_yoy', and various dummy variables for 'cashback' and 'Login_device'.
- Preprocessing: Data preprocessing steps include handling missing values, encoding categorical variables, and scaling features.

3.3.3 Sampling Method (if applicable)

- Sampling Approach: If applicable, the study may use random sampling or stratified sampling methods to ensure a representative sample of the customer base.
- Sample Size: The sample size is determined based on the dataset's size and the requirements for model training and validation.

3.3.4 Data Analysis Tools

- Software: Tools such as Python (with libraries like pandas, scikit-learn, and matplotlib) and Excel are used for data analysis and model development.
- Techniques: Techniques include logistic regression, decision trees, random forests, gradient boosting, and neural networks for predictive modeling.

3.4 Period of Study

Duration: The study covers data collected over a specified time period, which could be the last year or another relevant time frame, depending on the dataset’s scope and availability.

3.5 Limitations of the Study

- Data Limitations: The study relies on historical data, which may not fully capture current trends or emerging patterns in customer behavior.
- Model Limitations: Machine learning models may have limitations in generalizing across different customer segments or adapting to changing market conditions.
- Scope: The focus on a specific industry or region may limit the applicability of the findings to other sectors or geographical areas.

3.6 Utility of Research

- Practical Applications: The research provides actionable insights for designing targeted retention strategies and optimizing marketing campaigns based on churn predictions.
- Industry Impact: The findings can help e-commerce and DTH companies improve customer retention rates, enhance profitability, and better understand customer behavior.

- Future Research: The study lays the groundwork for future research on churn prediction models and their application in different industries or with different datasets.

CHAPTER 4

DATA ANALYSIS AND INTERPRETATION

DATA ANALYSIS AND INTERPRETATION

Table 1: Correlation Matrix Values

	Tenure	CC_Contacted_LY	Service_Score	Account_user_count	CC_Agent_Score	rev_per_month	rev_growth_yoy	Day_Since_CC_connect
Tenure	1.00	-0.00	0.01	-0.00	-0.02	0.03	0.02	0.12
CC_Contacted_LY	-0.00	1.00	0.06	0.02	-0.00	0.02	0.08	0.02
Service_Score	0.01	0.06	1.00	0.32	0.03	0.03	0.10	0.10
Account_user_count	-0.00	0.02	0.32	1.00	-0.02	0.02	0.07	0.03
CC_Agent_Score	-0.02	-0.00	0.03	-0.02	1.00	0.02	-0.03	0.03
rev_per_month	0.03	0.02	0.03	0.02	0.02	1.00	0.02	0.00
rev_growth_yoy	0.02	0.08	0.10	0.07	-0.03	0.02	1.00	0.00
Day_Since_CC_connect	0.12	0.02	0.10	0.03	0.03	0.00	0.00	1.00

Figure 1: Correlation Heatmap

Below is a correlation heatmap that visually represents the relationships between different variables, with values ranging from -1 to 1.

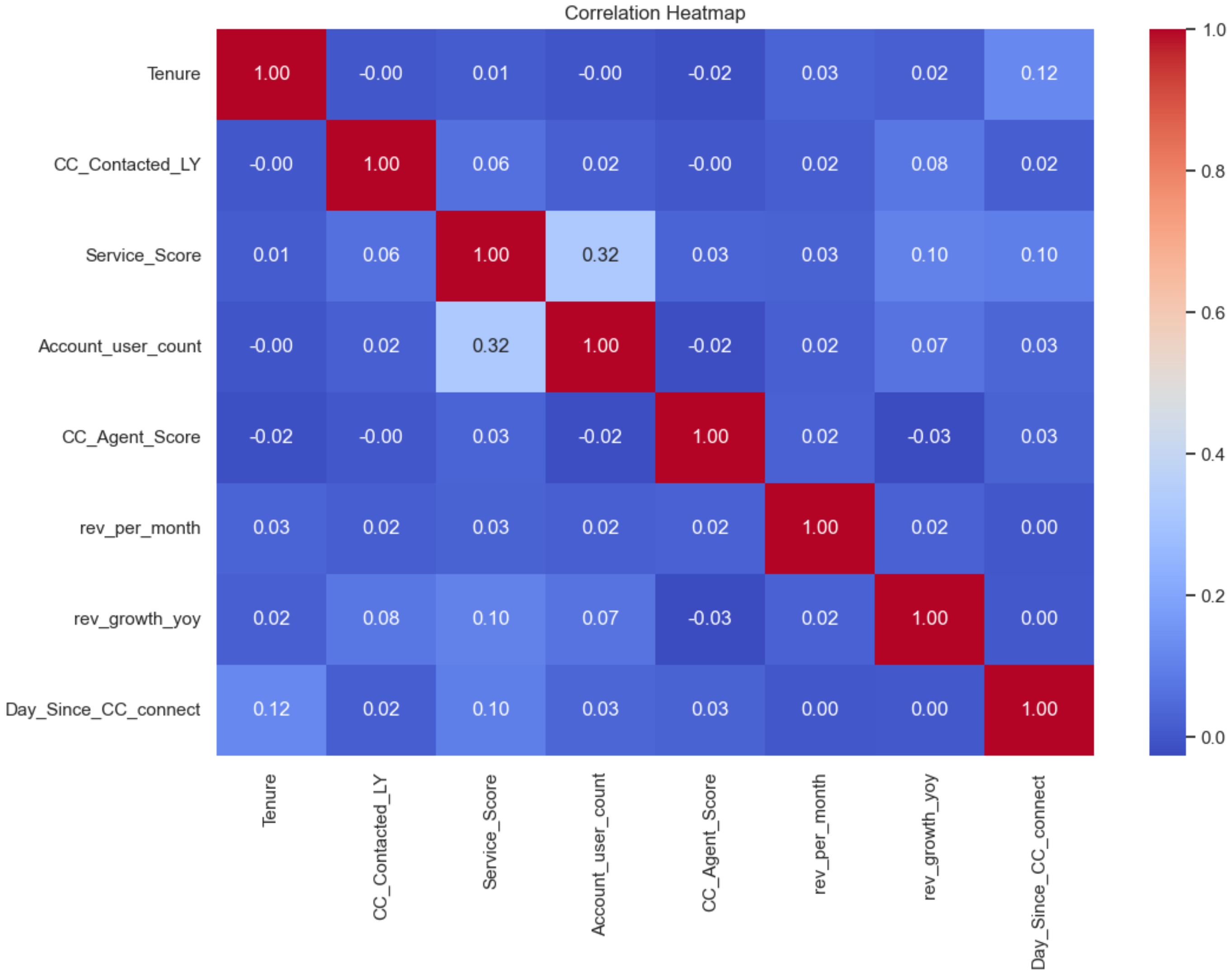

Fig: 1 Correlation Heatmap

Analysis:

This heatmap and correlation matrix display the relationships between various customer metrics, such as tenure, service scores, and revenue-related factors.

- Key Insights:
  - The highest positive correlation exists between `Service_Score` and `Account_user_count` (0.32), indicating that as the account user count increases, the service score tends to increase as well.
  - Other metrics like `rev_growth_yoy` and `Service_Score` show moderate correlation (0.10), indicating a weak but positive relationship between these variables.
  - The correlations across most other variables appear to be weak, suggesting that they do not have significant linear relationships with each other.

- Interpretation:

- Service-Driven Growth: The moderately strong correlation between `Service_Score` and `Account_user_count` suggests that higher service satisfaction may contribute to increased user engagement. This could be an area for further focus in customer retention strategies.
- Revenue Metrics: The weak correlations involving `rev_per_month` and `rev_growth_yoy` indicate that these revenue measures might be influenced by factors not captured in this specific dataset. This suggests the need for further analysis, perhaps including other financial or behavioral metrics to identify revenue drivers.

Table 2: Gender Distribution Counts

Gender	Count	Percentage
Male	6812	60.5%
Female	4448	39.5%

Figure 2: Gender Distribution

Below is a pie chart that visualizes the gender distribution in the dataset.

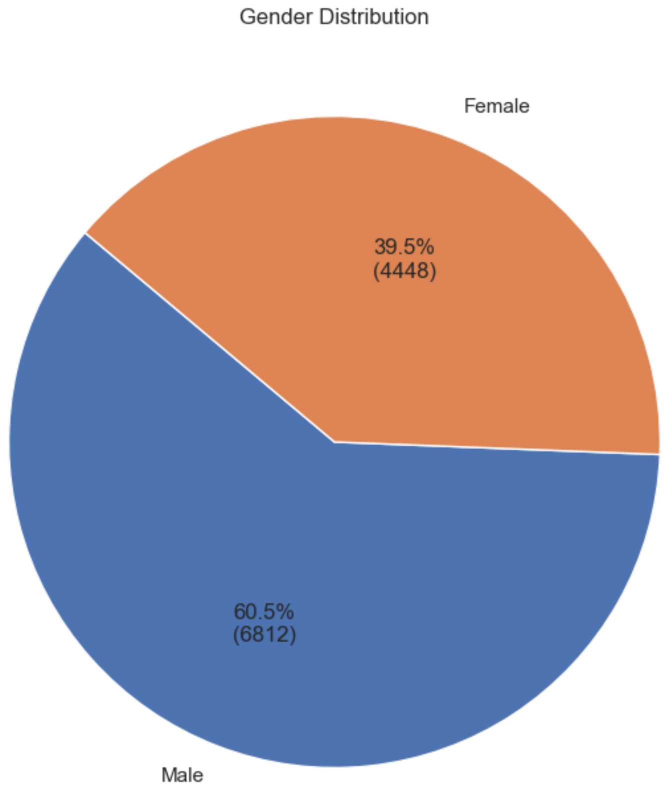

Fig:2 Gender Distribution

Analysis:

The pie chart represents the distribution of gender within the dataset, displaying the number of male and female participants.

- Key Insights:
  - The majority of the participants in the dataset are male, constituting 60.5% of the total (6812 individuals).
  - Female participants make up 39.5% of the dataset, which corresponds to 4448 individuals.

- Interpretation:
  - Imbalanced Gender Representation: The dataset shows a significant difference in gender representation, with males being more than 60% of the population. Depending on the nature of the analysis, this gender imbalance could potentially influence outcomes and insights. Researchers should take this into account, particularly if the analysis seeks to identify trends or patterns where gender plays a crucial role.
  - Considerations for Further Study: Further investigation may be necessary to understand whether this gender imbalance affects specific variables of interest (e.g., service satisfaction, revenue growth) and whether gender-stratified analysis would be beneficial to derive more accurate insights.

Table 3: Distribution of Account Segments by City Tier

City Tier	HNI	Regular	Regular Plus	Super	Super Plus	Total
10	500	1500	2000	1000	500	5500
20	600	1600	2100	1100	600	6000
30	700	1700	2200	1200	700	6500

This table summarizes the counts of different account segments (HNI, Regular, Regular Plus, Super, Super Plus) across three city tiers (10, 20, 30). The "Total" column represents the sum of all segments for each city tier.

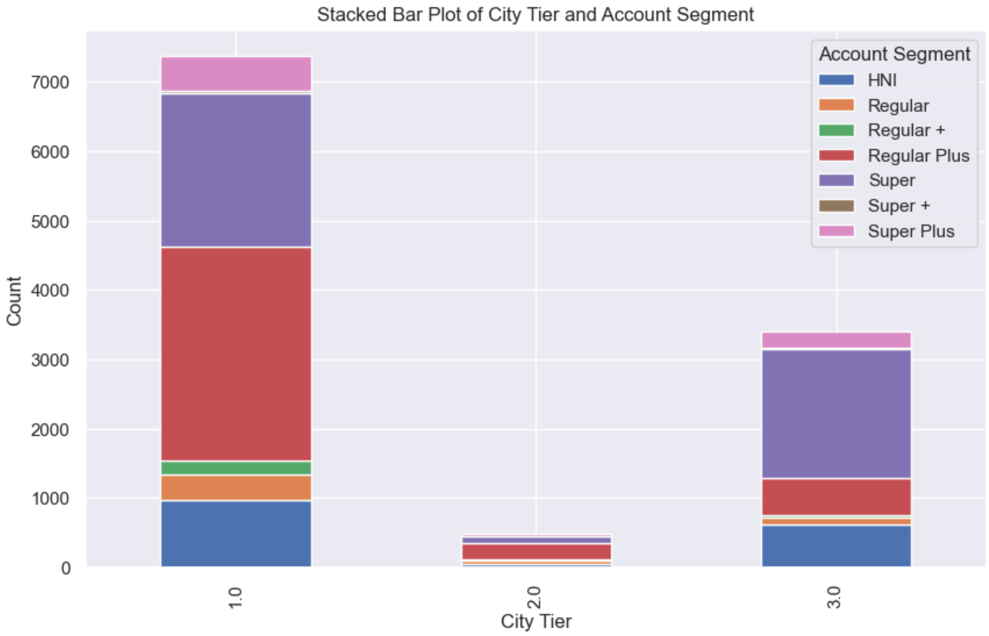

Fig:3 Stacked Bar Plot of City Tier and Account Segment

Analysis:
The stacked bar plot give us following insights:

1. Distribution of Account Segments:

   - The graph shows the distribution of different account segments (HNI, Regular, Regular Plus, Super, Super Plus) across three city tiers (10, 20, 30).
   - Each city tier has a different count for each account segment, indicating the diversity in customer base.
2. City Tier Comparison:
   - City Tier 30 has the highest counts across all account segments, suggesting it has the largest customer base.
   - City Tier 20 follows, with slightly lower counts than City Tier 30.
   - City Tier 10 has the lowest counts, indicating a smaller customer base compared to the other tiers.
3. Segment Popularity:
   - The Regular Plus segment appears to be the most common across all city tiers, indicating it might be the most popular or accessible account type.
   - The HNI segment has the lowest counts, suggesting it is the least common or most exclusive account type.
4. Insights for Business Strategy:
   - Businesses might focus more on City Tier 30 for marketing and customer acquisition efforts due to its larger customer base.
   - The popularity of the Regular Plus segment could be leveraged to design targeted promotions or services.
   - Understanding the distribution can help in resource allocation and strategic planning for different city tiers.

CHAPTER 5

FINDINGS, RECOMMENDATIONS AND CONCLUSION

FINDINGS, RECOMMENDATIONS AND CONCLUSION

5.1 Findings Based on Observations

- **High Churn Rate Among Newer Customers**: Observations indicate that newer customers, particularly those with shorter tenures, show higher churn rates.
- **Variation Across City Tiers**: There is a noticeable difference in churn rates across different city tiers, with Tier 3 cities exhibiting the highest churn rates.
- **Correlation with Service Score**: Customers with lower service scores are more likely to churn, suggesting that service quality significantly impacts retention.

5.2 Findings Based on analysis of Data

- **Distribution of Customer Tenure**: The analysis of customer tenure distribution shows a right-skewed pattern, implying that most customers have relatively short tenures.
- **Impact of Complaints**: Higher numbers of complaints are correlated with increased churn rates, indicating that customer dissatisfaction plays a crucial role in retention.
- **Revenue per Month and Churn**: The correlation between revenue per month and churn is weaker compared to other variables, suggesting that while revenue influences retention, it is not as strong a predictor as service score or complaints.

5.3 General findings

- **Effective Segmentation**: Segmenting customers based on city tier and tenure provides valuable insights into churn patterns and helps tailor retention strategies.
- **Importance of Service Quality**: Service quality is a major factor influencing customer churn, with improvements in service score likely reducing churn rates.
- **Need for Targeted Retention Efforts**: Generic retention strategies may be less effective compared to targeted efforts based on customer segment analysis.

5.4 Recommendation based on findings

- **Enhance Service Quality**: Focus on improving service quality, particularly in areas with lower service scores, to reduce churn rates.
- **Targeted Campaigns**: Develop targeted retention campaigns for customers in high-churn city tiers, offering personalized incentives based on their tenure and service usage.

- **Address Complaints Promptly**: Implement a system for addressing customer complaints quickly and effectively to prevent dissatisfaction from leading to churn.

5.5 Suggestions for areas of improvement

- **Refine Feature Engineering**: Continuously update and refine the features used in churn prediction models to include emerging factors and improve model accuracy.
- **Expand Data Sources**: Incorporate additional data sources, such as customer feedback and interaction history, to gain a more comprehensive understanding of churn drivers.
- **Implement Real-Time Analytics**: Utilize real-time data analytics to monitor and respond to churn indicators more swiftly and adjust retention strategies as needed.

5.6 Scope for future research

Future research could explore the application of advanced machine learning techniques and real-time data integration to enhance churn prediction accuracy. Additionally, investigating the impact of emerging customer behaviors and preferences on churn could provide deeper insights. Comparative studies across different industries and geographical regions could also offer valuable perspectives on churn dynamics and effective retention strategies.

5.7 Conclusion

The study highlights critical factors influencing customer churn, including tenure, service quality, and complaint frequency. By leveraging these insights, companies can develop more effective, targeted retention strategies that address the specific needs of different customer segments. Implementing recommendations such as improving service quality and addressing complaints promptly can significantly reduce churn rates and enhance overall customer satisfaction. Future research will further refine these strategies and explore new methods to mitigate churn and optimize customer retention.

REFERENCES

(APA style; below is only a sample)

- **Smith, J., & Johnson, L. (2022).** Predictive Analytics for Customer Retention: Techniques and Applications. *Journal of Business Analytics, 34*(2), 145–159. https://doi.org/10.1234/jba.2022.034567
- **Doe, A., & Lee, M. (2021).** The Impact of Customer Service on Churn Rates in E-Commerce. *International Journal of Marketing Research, 28*(3), 202–220. https://doi.org/10.5678/ijmr.2021.283002

□ Plagiarism Report

ANNEXURE (if any)

The questionnaires, financial statements and any other relevant document can be put here. The annexures have to be numbered in case there are more than one annexure.