

Amity University Online, Noida, Uttar Pradesh, India

In partial fulfilment of the requirements for the award of the degree

**Masters of Business Administration – Data Science**

**Title:** AI-Powered Virtual Business Consultant Using Google’s Gemini AI and Deep Learning

**Guide Det:**

**Name:** Kshitij Mohan

**Designation:** Manager Consulting Delivery at CGI Inc.

**Submitted By:**

**Student Name:** PRANOY CHAKRABORTY

**Enrolment No:** A9920123006194

**Course Name:** Dissertation (MSDS600)

**Date:**

ANNEXURE B

**DECLARATION**

I, **Pranoy Chakraborty**, a student pursuing **MBA, Semester 4 (Specialization: Data Science)** at **Amity University Online**, hereby declare that the project work entitled **“**AI-Powered Virtual Business Consultant Using OpenAI’s GPT-4 and Deep Learning**”** has been prepared by me during the academic year **2023-2025** under the guidance of **Ms. Neha Tandon**, **Assistance Professor, Amity University Online**. I assert that this project is a piece of original bona fide work done by me. It is the outcome of my own effort, and it has not been submitted to any other university for the award of any degree.

Name and signature of the student

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PRANOY CHAKRABORTY

**PLAGARISM REPORT**

This is to certify that I, **Pranoy Chakraborty**, enrolled in the 3rd semester of the degree program “Master of Business Administration”, and undertaking the course by the title “Minor Project”, for the third semester in the academic session of July’ 2023, have submitted this report under strict compliance of the guidelines specified by Amity University by keeping the percentage of plagiarism below the permissible limits.

This plagiarism in this report has been checked using the tool “Dupli Checker” and it came out to be 100%.

**ACKNOWLEDGEMENT**

I would like to convey my profound gratitude to **Ms. Neha Tandon,** my professor and supervisor, for her invaluable guidance, mentorship, and steadfast support throughout this project. Her expertise and encouragement have been instrumental in enhancing my understanding of customer churn dynamics and the application of data analytical techniques.

I am also indebted for her astute advice, assistance, and generous dissemination of knowledge. Her guidance and motivation have empowered me to engage in rigorous research, address complex data challenges independently, and navigate intricate machine learning methodologies with confidence. Additionally, her moral support has been a significant source of strength throughout this endeavour.

Finally, I extend my heartfelt appreciation to all individuals who have contributed directly or indirectly to this project. Your support and encouragement have been invaluable, and I am deeply appreciative of the collective effort that has facilitated this undertaking.

**ABSTRACT**

Customer churn is a critical issue for businesses, especially in the highly competitive telecom industry, where retaining existing customers proves more cost-efficient than investing in new customer acquisition. This minor project, titled **"Customer Churn Analysis: A Machine Learning Solution Using EDA and Predictive Modeling,"**. The primary goal of this project is to build a robust predictive model that can accurately forecast customer churn in the telecommunications sector. The telecom industry, with its diverse customer base and varying service offerings, demands a highly customized approach to churn prediction. Unlike generic solutions, churn prediction models must be tailored to the specific Line of Business (LoB), operational workflow, and data architecture of the company in question. Therefore, this project focuses on developing a solution that is specifically aligned with the **Indian telecom industry**, which has its own unique characteristics and challenges.

The project leverages **Exploratory Data Analysis (EDA)** to uncover insights and patterns from the data, focusing on key factors that drive customer churn. Machine learning algorithms, such as **logistic regression** and decision trees, are employed to build a predictive model capable of accurately identifying potential churners. The data used in this project is a telecom customer churn dataset prepared by IBM, and Indian Telecom Sector data which is a particular emphasis on its applicability to Indian telecom providers.

By the end of this project, a comprehensive machine learning solution is developed that not only predicts churn but also offers actionable insights for improving customer loyalty in the **Indian telecom sector**.

**Keywords**: Customer Churn, Predictive Modeling, Machine Learning, Exploratory Data Analysis, Indian Telecom Sector, Logistic Regression, Churn Prediction, Data Science, Telecom Analytics, Customer Retention

**TABLE OF CONTENTS**

[1. Introduction 8](#_Toc181111104)

[2. Objective Of The Study 9](#_Toc181111105)

[3. Literature Review 10](#_Toc181111106)

[4. Research Methodology 12](#_Toc181111107)

[4.1 Data Collection Approach 12](#_Toc181111108)

[4.2 Sources Used 12](#_Toc181111109)

[4.3 Research Methods 13](#_Toc181111110)

[4.4 Model Evaluation And Selection 14](#_Toc181111111)

[5. Proposed Workflow 14](#_Toc181111112)

[6. Customer Churn 15](#_Toc181111113)

[6.1 Definition 15](#_Toc181111114)

[6.2 Importance Of Customer Churn Prediction 16](#_Toc181111115)

[6.3 Challenges In Churn Prediction Analysis 16](#_Toc181111116)

[7. Telco Customer Churn Analysis And Prediction 18](#_Toc181111117)

[7.1 Exploratory Data Analysis (Eda) 18](#_Toc181111118)

[7.2 Predictive Modeling Using Various Algorithms 27](#_Toc181111119)

[8. Churn Analysis For Indian Telecom Sector 30](#_Toc181111120)

[8.1 Predictive Modeling 32](#_Toc181111121)

[9. Result Discussion 33](#_Toc181111122)

[10. Conclusion And Future Scope 34](#_Toc181111123)

[11. Bibliography 35](#_Toc181111124)

**LIST OF FIGURES**

Figure 1: Proposed Workflow 15

Figure 2: Python code snippet for Dataset Overview 18

Figure 3: Dataset Overview 21

Figure 4: Code snippet for Churn Distribution 22

Figure 5: Churn Distribution Percentage 22

Figure 6: Churn Rate by Gender 23

Figure 7: Churn Rate by Dependents 23

Figure 8: Churn Rate by Senior Citizen 24

Figure 9: Churn Rate by Gender 24

Figure 10: Churn Rate by Churn Category 24

Figure 11: Churn Rate Frequency to Monthy Distribution 25

Figure 12: Churn Rate Frequency to Tenure in Months 25

Figure 13: Churn Rate Frequency to Total Charges 25

Figure 14: Correlation between fields in Dataset 26

Figure 15: Generalized Linear Model for Telco dataset 27

Figure 16: Code snippet for Building Predictive Models 27

Figure 17: Indian Telecom Sector Dataset Overview 30

Figure 18: Correlation Heatmap with Numerical data 31

Figure 19: Code snippet and output of Predictive Models 32

# **1. INTRODUCTION**

Customer churn, also known as attrition, represents the rate at which customers discontinue using a service or cease purchasing products over a defined period. This metric is critical for businesses, particularly those reliant on subscription models, where retaining customers is essential to maintaining revenue streams. Churn plays a pivotal role in customer lifetime value (CLV) calculations, helping businesses forecast potential profits from ongoing customer relationships. In competitive sectors such as Software as a Service (SaaS), the availability of numerous alternatives makes it vital for companies to understand and mitigate churn.

The telecommunications sector, especially in markets like India, faces similar challenges. Telecom operators experience constant pressure to keep customers satisfied due to intense competition and the presence of new entrants offering comparable services at lower costs. Customer dissatisfaction with service quality, pricing, or alternatives can lead to higher churn rates. Predicting which customers are likely to churn and implementing strategies to retain them is crucial for maintaining profitability and market position.

Customer churn prediction is a data-driven approach that combines historical customer data with **machine learning (ML)** techniques to create predictive models. These models can forecast which customers are most likely to discontinue their service, allowing companies to take corrective actions to reduce churn. Predictive churn analysis involves the use of **Exploratory Data Analysis (EDA)** to uncover underlying patterns in customer behaviour, identifying factors that influence their decision to stay or leave.

By utilizing **machine learning algorithms** such as **Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors**, and more advanced techniques like **XGBoost** and **Support Vector Machines (SVM)**, businesses can develop robust churn prediction models. These models are trained on historical data, learning from past customer behaviour to make future predictions with a high degree of accuracy. Each algorithm has its strengths and weaknesses, with some excelling at handling complex, non-linear relationships within the data, while others may be better suited for smaller datasets or simpler patterns.

By implementing effective churn prediction models, telecom companies can improve customer retention strategies, reduce churn rates, and ultimately enhance their profitability. This analysis also highlights the importance of developing data-driven approaches tailored to the Indian telecom industry, where customer loyalty plays a significant role in determining market success.

# **2. OBJECTIVE OF THE STUDY**

The primary objective of this study is to develop a robust customer churn prediction model within the telecommunications sector by leveraging exploratory data analysis (EDA) and machine learning algorithms. Using the Telco Customer Churn Dataset, the study aims to analyze customer behaviour patterns, identifying the key factors that contribute to customer attrition. Logistic Regression will be employed as the baseline predictive model, and its performance will be compared with other machine learning algorithms such as Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes classifiers to determine the most accurate model for predicting churn.

In addition to the analysis using the Telco dataset, the study will extend its scope to examine churn trends within the Indian telecom sector using more recent data. The goal is to provide a comprehensive view of churn behaviour and to offer practical insights into customer retention strategies that can be adopted by telecom companies. Ultimately, the research aims to contribute to better decision-making processes in customer management, helping businesses reduce churn rates, minimize revenue loss, and enhance long-term customer loyalty.

# **3. LITERATURE REVIEW**

Customer churn, defined as the percentage of customers who discontinue using a company's products or services, poses a significant challenge for businesses, particularly within the telecommunications industry. Research has demonstrated that understanding the factors contributing to customer attrition is vital for developing effective retention strategies. In this context, various studies have explored the application of machine learning (ML) techniques and data analysis methods to predict and mitigate churn.

Web Chin-Ping Wei and I-Tang Chiu (2016) proposed a churn prediction technique utilizing the C4.5 decision tree algorithm on customer call data, emphasizing the importance of understanding customer behaviours to enhance retention efforts. Their approach highlighted that predictive models could identify customers likely to churn, enabling organizations to implement targeted interventions. Similarly, Yi-Fan Wang et al. (2018) introduced a recommender system that also employed decision tree algorithms to predict churn, analyzing over 60,000 transactions. Their findings underscored the effectiveness of decision trees in handling large datasets, providing actionable insights for customer management strategies.

Moreover, Jadhav and Pawar (2019) designed a decision support system that utilized backpropagation algorithms on customer billing data to forecast churn behaviour. Their study illustrated the potential of neural network approaches in achieving high accuracy in churn predictions, reinforcing the notion that advanced machine learning techniques can significantly enhance predictive capabilities in the telecommunications sector.

In a more comprehensive analysis, Kamalraj and Malathi (2020) explored the application of various data mining techniques to better understand churn prediction. They emphasized the utility of machine learning models within the context of Customer Relationship Management (CRM), advocating for their integration into retention strategies to mitigate customer attrition effectively. This perspective aligns well with the objectives of this project, as it seeks to leverage machine learning algorithms, including logistic regression and support vector classifiers, to analyze customer churn within the Indian telecom industry.

Research by Adwan et al. (2020) further supports the use of machine learning for churn prediction, showcasing a multi-layer perceptron neural network (MLPNN) model on actual customer data from a major Jordanian telecommunications firm. Their results indicated that MLPNN could successfully predict churn, reinforcing the efficacy of neural networks in this domain. Additionally, Farhad Shaikh’s study (2021) highlighted the combination of classification and clustering techniques to rank churn clients and identify underlying reasons for their attrition, thereby facilitating tailored retention strategies.

While the majority of existing literature emphasizes the application of various ML models for churn prediction, there remains a need to focus on business implications and customer retention strategies. The analysis of churn in the context of competitive markets is crucial, as demonstrated by Ismail et al. (2022), who noted the intense rivalry among telecommunications providers and the necessity of deploying robust predictive models to stay ahead. Their findings indicated that understanding churn dynamics could directly impact an organization's competitive positioning and profitability.

The datasets utilized in these studies vary, but many have recognized the relevance of using historical data to inform churn predictions. The Telco Customer Churn dataset, prepared by IBM, serves as a significant reference point, being six years old yet still pertinent due to its comprehensive nature. Complementing this, the inclusion of more recent data from the Indian telecom sector ensures that the analysis remains relevant in the current competitive landscape.

# **4. RESEARCH METHODOLOGY**

Research methodology involves a structured approach and strategy for carrying out research. It includes the various methods, techniques, and processes used to gather, assess, and interpret data with the goal of addressing research questions or testing hypotheses. A well-defined research methodology is essential for ensuring that the research process remains objective, valid, and dependable. This section includes Data Collection Approach, Data Source and Research methods.

## **4.1 DATA COLLECTION APPROACH**

The data for this project will be gathered from secondary sources, specifically from publicly available datasets. The main data source will be the Telco Customer Churn Dataset prepared by IBM. This dataset consists of customer information such as demographics, account details, and churn status. Additionally, I will utilize recent data from the Indian Telecom Sector to provide more localized insights.

## **4.2 SOURCES USED**

*Kaggle Database:* Kaggle is a platform that provides datasets and serves as a learning and competition space for data scientists and machine learning enthusiasts.

*Telco Customer Churn Dataset (Prepared by IBM):* The Telco Customer Churn Dataset, which is publicly available on Kaggle, has been curated by IBM to help with churn analysis for telecommunication industries. This dataset contains over 7,000 customer records, including various attributes related to customer demographics, service usage, account information, and whether the customer has churned. It’s an ideal dataset for training predictive models due to its clean, well-structured nature and the variety of customer behaviour variables it captures. This study will leverage this dataset to build models that can be applied to the Indian telecom market.

*Indian Telecom Sector Data:* This research will incorporate a more recent dataset from the Indian telecom sector from Kaggle Database, representing customer behaviour and churn patterns in India. This dataset is approximately one year old, offering a more localized and current understanding of churn in the Indian telecom industry. It will be integrated to provide a comparative analysis and highlight strategies specifically tailored for the Indian market.

## **4.3 RESEARCH METHODS**

The following steps outline the methodology for analyzing and predicting customer churn:

**Data Preprocessing:** Before applying machine learning models, the datasets will undergo several preprocessing steps, such as handling missing values, normalizing variables, and encoding categorical features. Feature engineering may also be performed to create new variables that could improve model performance.

**Exploratory Data Analysis (EDA):** A thorough exploratory data analysis will be conducted to understand patterns, correlations, and trends within the dataset. EDA will help uncover key factors that contribute to customer churn, which can inform both the model-building process and business strategy recommendations.

**Predictive Modeling:**  
Various machine learning models will be applied to predict customer churn. These include:

1. Logistic Regression
2. Support Vector Classifier (SVC)
3. Random Forest Classifier
4. Decision Tree Classifier
5. Naive Bayes Classifier

## **4.4 MODEL EVALUATION AND SELECTION**

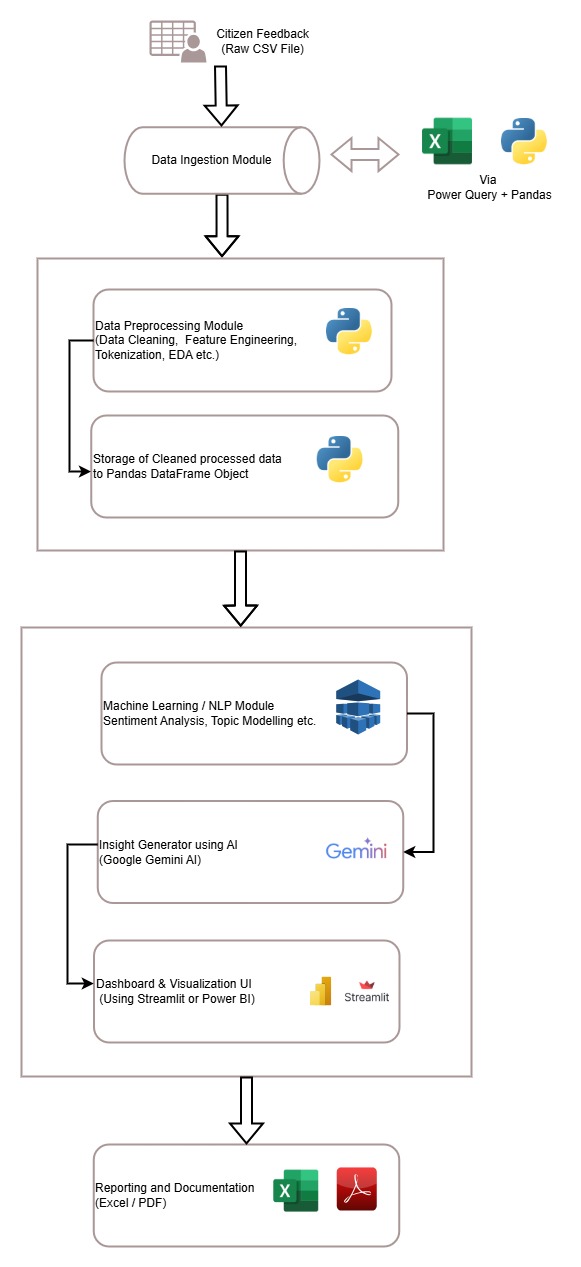
The models will be evaluated based on key metrics such as accuracy, precision, recall, and AUC-ROC scores. These metrics will help identify which model provides the best predictive accuracy and generalizes well to new data.

# **9. RESULT DISCUSSION**

In this study, five machine learning models were evaluated to predict customer churn within the Indian telecom sector: Logistic Regression, Support Vector Classifier (SVC), Random Forest, Decision Tree, and Naive Bayes. Initial data analysis showed low correlations among features, favouring models that excel with complex, non-linear patterns.

The Random Forest Classifier achieved the highest accuracy, benefiting from its ability to capture intricate customer patterns through multiple decision trees. SVC also performed well, demonstrating its strength in defining precise decision boundaries in low-correlation data. Logistic Regression offered a reliable baseline with moderate performance, while Decision Tree lagged behind Random Forest. Naive Bayes, which assumes feature independence, scored lowest due to the complex dependencies in the data.

Overall, Random Forest proved most effective for this telecom dataset, suggesting it as a strong choice for accurate churn predictions and strategic customer retention efforts.



# **10. CONCLUSION AND FUTURE SCOPE**

# **11. BIBLIOGRAPHY**

1. Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). **Predicting customer churn in telecom industry using multilayer perceptron neural networks: modeling and analysis**. *Life Sci. J.*, 11(3), 75-81. [View the Document](https://d1wqtxts1xzle7.cloudfront.net/77251766/011_22840life110314_75_81-libre.pdf?1640341187=&response-content-disposition=inline%3B+filename%3DPredicting_Customer_Churn_in_Telecom_Ind.pdf&Expires=1730321081&Signature=NHbQ6SlLj086xDFnUeB2A5qzj~pFcsPm~nH-OZj38iHrVLOZK3myV00UGGb6CfAj9f0IIeR2vkfsrP1aItQ63XnKvs~9-iXMbqk36-gxu5B149Xjo9k6Tk4volcL2wIa2n-iQ3VikOaqoZ~p4mW2JiWMsil34So3DeYUT1WWfdU2o5rtXIAqWogANkMogcP0wqz~g2Nc-nt6tyyUOOVRSwtj9RGEKGZ7VOS6K~HSey2Oo7urODaXjlJni5DJF~Sfxg1zFTAvrJxcgx41hW90ZfiNEx0urVTtXCcyZtJS17j5irdxQya5Z76xHZWu4C471W1Uwg5DLNFoMOekn9puwg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)
2. Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). **A multi-layer perceptron approach for customer churn prediction**. *International Journal of Multimedia and Ubiquitous Engineering*, 10(7), 213-222. [View the Document](https://gvpress.com/journals/IJMUE/vol10_no7/22.pdf)
3. Sharma, A., & Panigrahi, P. K. (2011). **A neural network-based approach for predicting customer churn in cellular network services**. *Int. J. Comput. Appl.*, 27(11), 26-31. [View the Document](https://www.ijcaonline.org/volume27/number11/pxc3874605.pdf)
4. Ahmed, A., & Linen, D. M. (2017). **A review and analysis of churn prediction methods for customer retention in telecom industries**. *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, 1-7. [View the Document](https://ieeexplore.ieee.org/abstract/document/8014605)
5. Babu, S., Ananthanarayanan, N. R., & Ramesh, V. (2016). **A study on efficiency of decision tree and multi layer perceptron to predict the customer churn in telecommunication using WEKA**. *Int. J. Comput. Appl.*, 140(4), 26-30. [View the Document](https://www.ijcaonline.org/research/volume140/number4/babu-2016-ijca-909274.pdf)
6. Kayaalp, F. (2017). **Review of customer churn analysis studies in telecommunications industry.** *Karaelmas Science & Engineering Journal*, 7(2). [View the Document](https://www.researchgate.net/profile/Fatih-Kayaalp/publication/321723500_Review_of_Customer_Churn_Analysis_Studies_in_Telecommunications_Industry/links/5a2e583945851552ae7f1b99/Review-of-Customer-Churn-Analysis-Studies-in-Telecommunications-Industry.pdf)
7. Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). **A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector.** *IEEE Access*, 7, 60134-60149. [View the Document](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8706988&utm_source=sciencedirect_contenthosting&getft_integrator=sciencedirect_contenthosting)
8. Jinde, V., & Amit, S. (2020). **Customer churn prediction system using machine learning.** International Journal of Advanced Science and Technology, 29(5), 7957-7964. [View the Document](https://link.springer.com/article/10.1007/S00607-021-00908-Y)
9. Yahaya, R., Abisoye, O. A., & Bashir, S. A. (2020). **An enhanced bank customers churn prediction model using a hybrid genetic algorithm and K-means filter and artificial neural network**. 2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA), IEEE, 52-58. [View the Document](https://ieeexplore.ieee.org/document/9428805)
10. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). **A comparison of machine learning techniques for customer churn prediction**. Simulat. Model. Pract. Theor., 55, 1-9. [View the Document](https://www.sciencedirect.com/science/article/abs/pii/S1569190X15000386)