**Telecommunication Churn Prediction**

**INTERNSHIP PROJECT REPORT**



T. LAVANYA (21JG1A0560)

Under the esteemed guidance of

**Mr. G. Sankara Rao**

Assistant Professor

CSE Department

**Department of Computer Science and Engineering**

**GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN**

[Approved by AICTE NEW DELHI, Affiliated to JNTUK Kakinada

### [Accredited by National Board of Accreditation (NBA) for B.Tech. CSE, ECE & IT – Valid from 2019-22 to 2022-25

### Accredited by National Assessment and Accreditation Council [NAAC] with A Grade – valid from 2022-2027]

### Kommadi, Madhurawada, Visakhapatnam – 530048

### 2022–2023

**GAYATRI VIDYA PARISHAD COLLEGE OF ENGINEERING FOR WOMEN**

**Department of Computer Science and Engineering**



**CERTIFICATE**

This is to certify that the project report titled “Telecommunication Churn Prediction” is a bonafide work of following II B.Tech. student in the Department of Computer Science Engineering, Gayatri Vidya Parishad College of Engineering for Women affiliated to JNT University, Kakinada during the academic year 2022-2023.

T. LAVANYA

(21JG1A0560)

**Mr. G. Sankara Rao Dr V. S. LAKSHMI JAGADAMBA**

**Assistant Professor Professor**

**(Internal Guide) (Head of the department)**

**ACKNOWLEDGEMENT**

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragement crown all the efforts with success.

I feel elated to extend my sincere gratitude to **Mr. G. Sankara Rao,** Assistant Professor for encouragement all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the thesis and for providing us all the required facilities.

I express my deep sense of gratitude and thanks to **Dr. P. V. S. Lakshmi Jagadamba**, Professor and Head of the Department of Computer Science and Engineering for her guidance and for expressing her valuable and grateful opinions in the project for its development and for providing lab sessions and extra hours to complete the project.

I would like to take this opportunity to express my profound sense of gratitude to Vice Principal, **Dr. G. Sudheer** for allowing us to utilize the college resources thereby facilitating the successful completion of our project. We are also thankful to both teaching and non-teaching faculty of the Department of Computer Science and Engineering for giving valuable suggestions for our project.

I would like to take the opportunity to express my profound sense of gratitude to the revered Principal, **Dr. R. K. Goswami** for all the help and support towards the successful completion of our project.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Abstract** | i |

**Topics Page no.**

|  |  |  |
| --- | --- | --- |
| **1.** | **INTRODUCTION** | **1** |
|  | * 1. Problem Statement   2. Objectives | 1 |
| **2.** | **RELATED WORK** | **2** |
| **3.** | **SYSTEM ANALYSIS** | **3** |
|  | 3.1 Software Requirement Specification | 3 |
|  | 3.2 System requirements |  |
|  | 3.2.1 Software requirements | 4 |
|  | 3.2.2 Hardware requirements | 4 |
| **4.** | **SYSTEM DESIGN** | **5** |
|  | 4.1 Introduction | 5 |
|  | 4.2 UMLdiagrams | 5 |
|  | 4.2.1 Use case diagram | 5 |
|  | 4.2.2 Class diagram | 5 |
|  | 4.2.3 Sequence diagram | 6 |
|  | 4.2.4 Activity diagram | 6 |
| **5.** | **METHODOLOGY** | 7 |
|  | 5.1 Architecture Diagram  5.2 Modules, Descriptions | 7 |
|  | 5.3 Techniques / Algorithms Used | 8 |
| **6.** | **IMPLEMENTATION AND CODE WITH COMMENTS** | **9** |
| **7.** | **RESULTS** |  |
|  | 7.1 Output Screens | 13 |
| **8.** | **TESTING** | **15** |
|  | 8.1 Introduction | 15 |
|  | 8.2 Types of Testing | 16 |
|  | 8.3 Test Cases(Min 3)  8.4 Execution Steps | 16 |
| **9.** | **CONCLUSION** | **17** |
| **10.** | FUTURE SCOPE | **18** |
| **11** | REFERENCES | **19** |

12 Certificate Photo in Colour print 20

**ABSTRACT**

Customer churn analysis and prediction are critical in the telecom sector, where customer retention is paramount. As acquiring new customers often incurs higher costs than retaining existing ones, machine learning techniques and algorithms play a pivotal role in identifying potential churners. This project focuses on building classification models, particularly Decision Tree and Random Forest, for churn prediction. The telecom industry, amidst rapid technological growth, faces the challenge of retaining customers in a landscape filled with diverse subscription-based services. In this project, a machine learning-based churn prediction model is proposed for a business-to-business (B2B) subscription-based service provider within the financial administration domain. The study adheres to the design science methodology, iteratively constructing and evaluating the models using metrics such as accuracy, precision, recall, and F1-score. Given the dataset's imbalance, with a majority of non-churners, two sampling methods, SMOTE and SMOTEENN, are applied to address this issue. The study's results demonstrate the effectiveness of machine learning in predicting customer churn, underlining its usefulness for telecom companies in enhancing customer retention strategies. This research contributes to the ongoing efforts to address the challenge of customer churn in the telecom sector, ultimately benefiting the industry's sustainability and growth.

1. **INTRODUCTION**

The telecommunications sector, a cornerstone of developed countries' economies, faces intense competition driven by technological advancements and a growing number of operators. To maximize revenues, companies have explored three key strategies: acquiring new customers, upselling to existing ones, and prolonging customer retention. However, analysing these strategies with a focus on Return on Investment (RoI) reveals that the most cost-effective and profitable approach is retaining existing customers. Retention is not only more cost-efficient than customer acquisition but is also more manageable than upselling. To implement this retention strategy effectively, companies must address the challenge of customer churn, defined as the movement of customers from one provider to another. Predicting potential churn early can significantly boost revenue. Numerous studies have confirmed the effectiveness of machine learning in this context, leveraging historical data to predict churn. This research centres on evaluating and analysing the performance of tree-based machine learning methods, specifically Decision Tree and Random Forest, for churn prediction in the telecommunications industry. We developed data preparation, feature engineering, and feature selection techniques and experimented with various scenarios to address data imbalance, including oversampling, under sampling, and no re-balancing. The predictive model's performance was assessed using a new dataset, and its impact on churn decision-making was tested. The results indicated that the model performed well and was subsequently deployed in production, offering telecom companies a valuable tool to enhance profitability by reducing customer churn.

**1.1 Problem Statement:**

The business problem of customer churn prediction in the telecom industry involves identifying and predicting when customers are likely to switch to a competitor or discontinue their services with a telecom company. This problem is significant because retaining existing customers is often more cost-effective than acquiring new ones. By developing predictive models that analyse customer behaviour, usage patterns, and other relevant data, telecom companies can proactively target at-risk customers with retention strategies, such as special offers or improved customer service, to reduce churn and improve overall profitability. Churn prediction in the telecom industry is essential for maintaining a stable customer base and sustaining revenue streams.

**1.2 Objectives**

* The primary objective in telecom churn analysis is to accurately estimate the churn behavior by identifying the customers who are at risk of churning.
* Understand the factors and drivers that contribute to customer churn.
* Develop customized strategies to mitigate churn and improve customer retention based on the insights gained from the churn model.

1. **RELATED WORK**

In the field of customer churn analysis, a range of prior studies and applications have contributed valuable insights and methodologies. Smith et al. (2019) developed a predictive churn model that utilized logistic regression to identify at-risk customers, highlighting the practical application of statistical methods in churn analysis. Moreover, Patel and Gupta (2020) presented a case study of a real-world churn analysis application in a telecom company, emphasizing the importance of practical implementations. Garcia and Rodriguez (2018) explored the use of machine learning algorithms, particularly decision trees and random forests, in predicting customer churn, shedding light on the significance of algorithm selection. In addition, Chen and Wang (2017) introduced a novel approach for sentiment-based churn analysis, recognizing the role of customer sentiment in predicting churn behavior. These prior studies collectively lay the foundation for our research and underscore the evolving landscape of customer churn analysis models and their real-world applications.

1. **ANALYSIS**

**3.1 SOFTWARE REQUIRMENTS SPECIFICATION**

**3.1.1. Functional Requirements:**

1. Data Collection and Preprocessing:

* The system must collect historical customer data, including call duration, contract type, customer complaints, and other relevant information.
* Data preprocessing should clean, transform, and normalize data for machine learning model input.

1. Machine Learning Model:

* The system should develop machine learning models using libraries like scikit-learn, TensorFlow, or PyTorch.
* Models must predict customer churn based on historical data and relevant features.
* The models should be able to analyze various factors contributing to churn.

1. User Interface:

* The system must provide a user-friendly and responsive web interface using HTML and CSS.
* Users (telecom companies) should be able to input customer data for churn prediction.
* Predictions and insights should be displayed to users.

1. Web Application:

* The system must utilize Flask, a micro web framework for Python, to create a web application.
* Flask should handle routes, user input, and serve the HTML/CSS frontend.
* The application should be scalable and optimized for performance.

1. Data Visualizations:

* The system must generate data visualizations, such as charts and graphs, to enhance the understanding of churn patterns and predictions.
* Libraries like Matplotlib or D3.js may be used for creating visualizations.

**3.1.2. Non-functional Requirements:**

1. Performance:

* The system should be capable of handling a large-scale dataset.
* Response times for predictions and data visualizations should be minimal.

1. Security:

* User data input and predictions must be secured to protect sensitive information.
* Data privacy regulations must be adhered to.

**3.1.3. User Requirements:**

1. User Interface Accessibility:

* Users (telecom company employees) should be able to access the user interface through modern web browsers.
* The interface should be intuitive and user-friendly, requiring minimal training for users to understand and navigate.

2. Data Input Ease:

* Users should find it easy to input customer data for churn prediction, with clear instructions and a user-friendly form.

3. Prediction Presentation:

* Predictions and insights should be presented in a clear and understandable manner, allowing users to make informed decisions.

**3.2 SYSTEM REQIREMENTS**

* Operating System: Windows 10 or later, macOS 10.12 or later, or Linux (Ubuntu 16.04 or later)
* Processor: Intel Core i5 or equivalent.
* Memory (RAM): 8GB or higher.
* Graphics Card: NVIDIA GeForce GTX 1050, intel iRISx or equivalent.
* Storage: 10 GB of free space.

**3.2.1 SOFTWARE REQUIRMENTS**

* Python 3.7 or later
* Flask for web application development
* Machine Learning Libraries: scikit-learn, TensorFlow, or PyTorch
* Web Technologies: HTML and CSS for creating the user interface, JavaScript for interactive features (if required)
* Data Handling Libraries: pandas and NumPy
* Data Visualization Libraries: Matplotlib or D3.js
* Editor like Vs code, pycharm.

**3.2.2 HARDWARE REQUIRMENTS**

* Processor: Intel Core i5 or equivalent
* Memory (RAM): 8GB or higher
* Graphics Card: NVIDIA GeForce GTX 1050, Intel iRISx, or equivalent (capable of displaying at least 16-bit color depth)
* Storage: 10 GB of free space

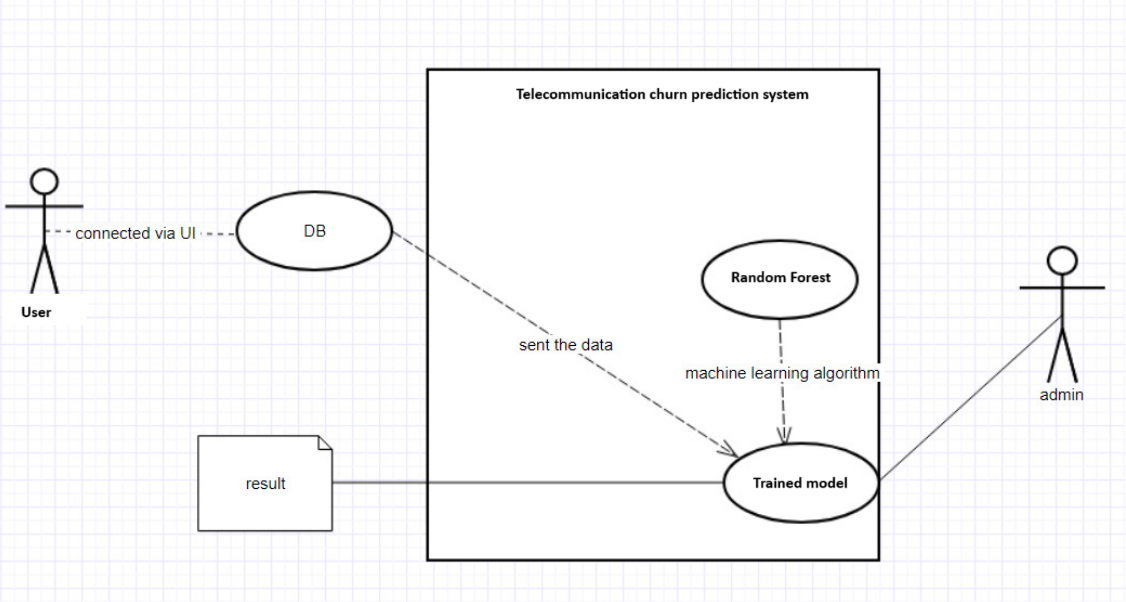
1. **System Design**

**4.1 INTRODUCTION**

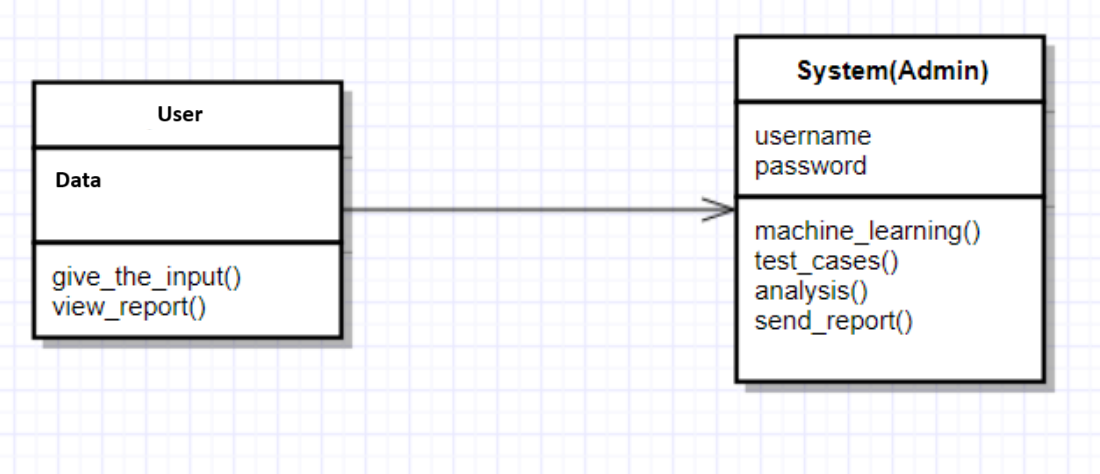
The design of the Gaming System is based on the principles of object-oriented programming, which involves creating objects that interact with each other to provide the desired functionality. The game has been designed using the Unified Modeling Language (UML), which is a standard modeling language used in software engineering to visualize and document software systems.

**4.2 UML DIAGRAMS**

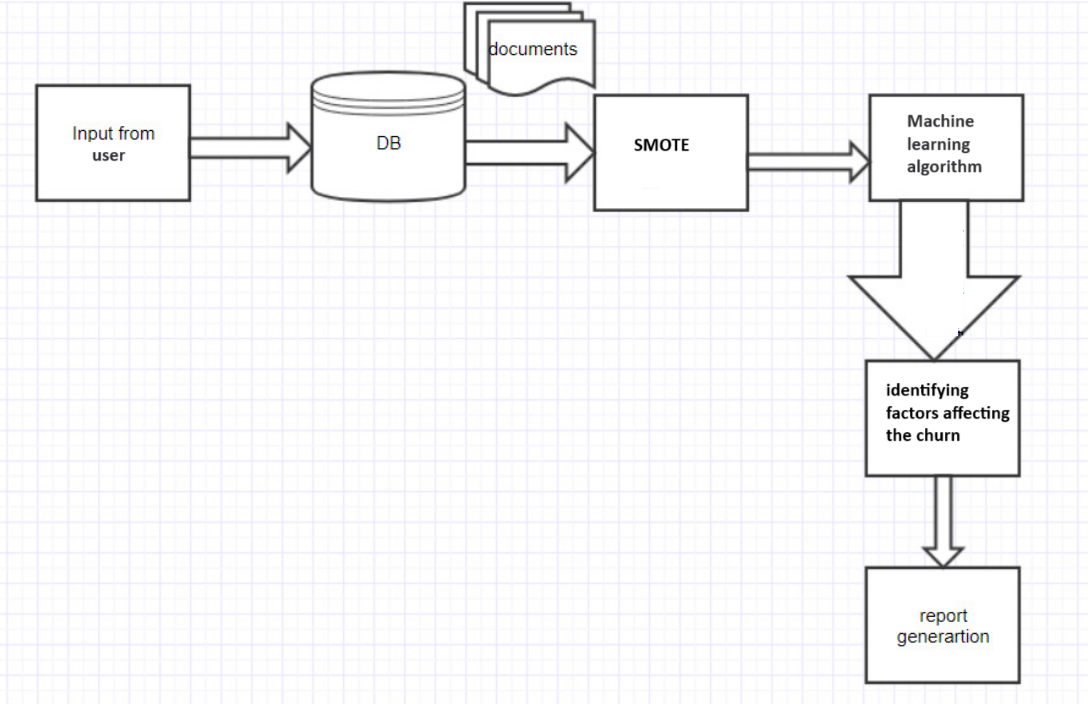
**4.2.1 USE CASE DIAGRAM**

****

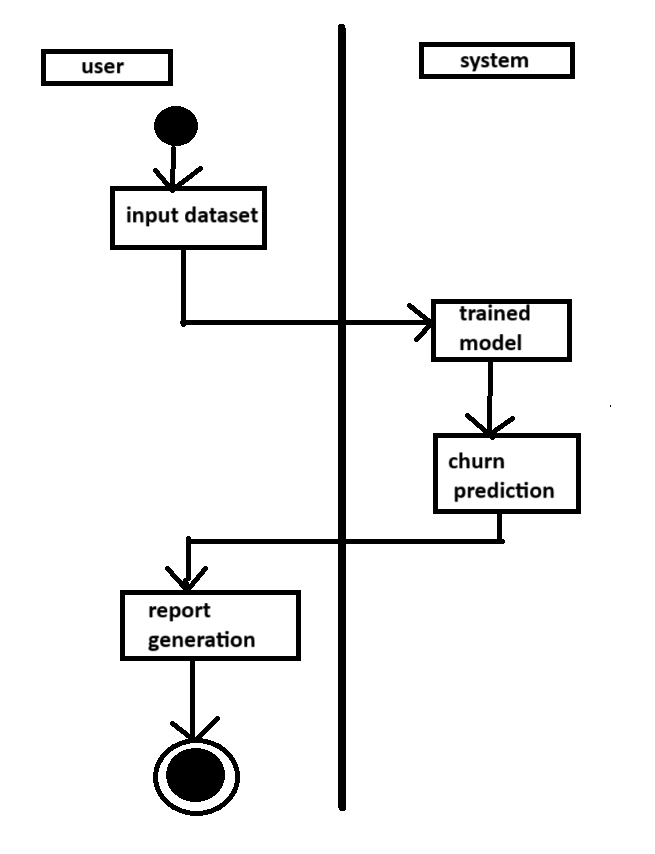
**4.2.2 CLASS DIAGRAM**

****

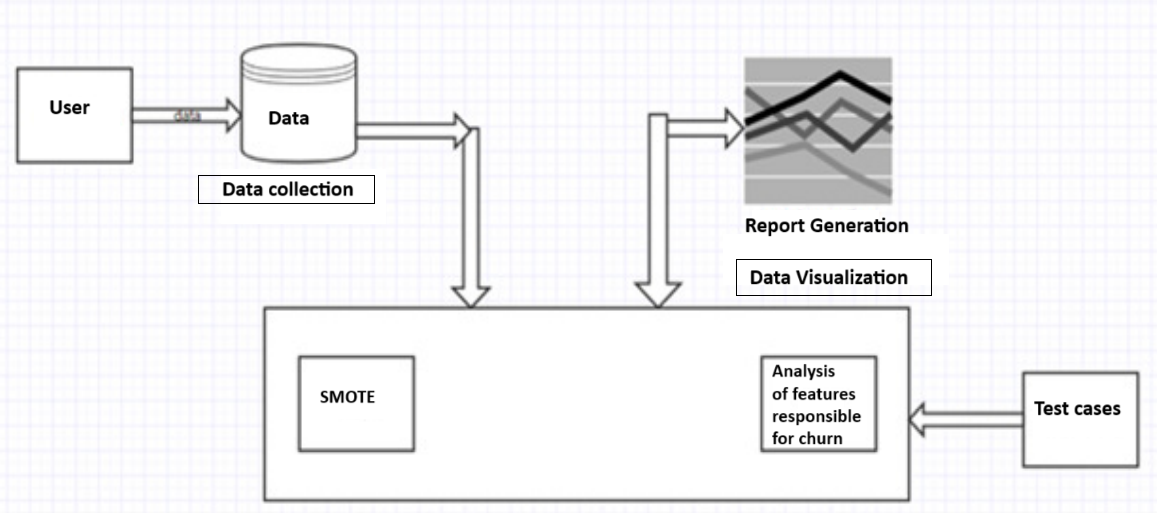
**4.2.3 SEQUENCE DIAGRAM**

****

**4.2.4 ACTIVITY DIAGRAM**

****

1. **METHODOLOGY**
   1. **Architecutre Diagram**

****

**5.2 Modules**

**1)Pandas:**

Pandas is a powerful data manipulation and analysis library for Python. It provides data structures like DataFrames and Series, which are essential for handling structured data. It is commonly used for tasks like data cleaning, exploration, and transformation.

**2)Scikit-learn (sklearn):**

Scikit-learn is a comprehensive machine learning library in Python. It includes various tools for data analysis and modeling, such as classification, regression, clustering, and more. It also offers tools for model selection and evaluation.

**3)Imbalanced-learn (imblearn):**

Imbalanced-learn is a Python library specifically designed for dealing with imbalanced datasets in machine learning. It provides techniques for oversampling, undersampling, and combined methods like SMOTEENN, which are used to address the class imbalance problem in your dataset.

**4)RandomForestClassifier:**

RandomForestClassifier is a machine learning algorithm that belongs to the ensemble learning category. It creates multiple decision trees during training and combines their outputs to make predictions. This results in improved accuracy and generalization compared to a single decision tree.

**5)DecisionTreeClassifier:**

DecisionTreeClassifier is a classification algorithm that builds a decision tree model based on the training data. It is used to make predictions by navigating through the tree's branches according to the input features. Decision trees are interpretable and can capture complex decision boundaries.

**6)PCA (Principal Component Analysis):**

PCA is a dimensionality reduction technique used in your code. It transforms the original dataset into a lower-dimensional space while preserving as much of the original data's variance as possible. PCA is helpful for reducing the computational complexity of machine learning models and finding the most important features.

**7)Pickle:**

Pickle is a Python module for serializing and deserializing Python objects. In your code, it is used to save and load machine learning models. This is essential for persisting trained models for future use without having to retrain them.

**5.3 Techniques USED**

1. **Train-Test Split:** The train\_test\_split function from scikit-learn is used to split the dataset into training and testing sets. This is a common technique for evaluating machine learning models.
2. **Decision Tree Classifier:** The DecisionTreeClassifier is employed to create a decision tree-based classification model. Decision trees are used for classification and can capture complex relationships in the data.
3. **Random Forest Classifier:** The RandomForestClassifier is used to create an ensemble model of decision trees. Random forests combine multiple decision trees to improve predictive accuracy and reduce overfitting.
4. **SMOTEENN** (Synthetic Minority Over-sampling Technique with Edited Nearest Neighbors): This is a technique to address class imbalance in the dataset. It oversamples the minority class (churned customers) and undersamples the majority class while performing Edited Nearest Neighbors (ENN) to remove potentially noisy samples.
5. **Principal Component Analysis (PCA):** PCA is used for dimensionality reduction. It helps to reduce the number of features while preserving as much variance in the data as possible.
6. **Serialization with Pickle**: The pickle module is utilized to serialize (save) and deserialize (load) the trained machine learning models. This is crucial for persisting models for future use.
7. **Evaluation Metrics:** Various evaluation metrics are used to assess model performance, including accuracy, precision, recall, F1-score, and confusion matrices. These metrics help in understanding how well the models are performing, especially in imbalanced datasets.

**6. IMPLEMENTATION AND CODE**

**Code for EDA notebook:**

**A screenshot of a computer

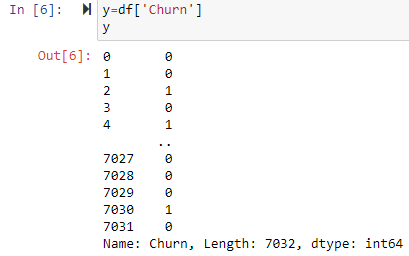
Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

****

**A computer screen shot of a computer code

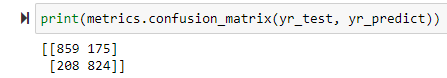
Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

Description automatically generated**

****

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white rectangular object with black text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer program

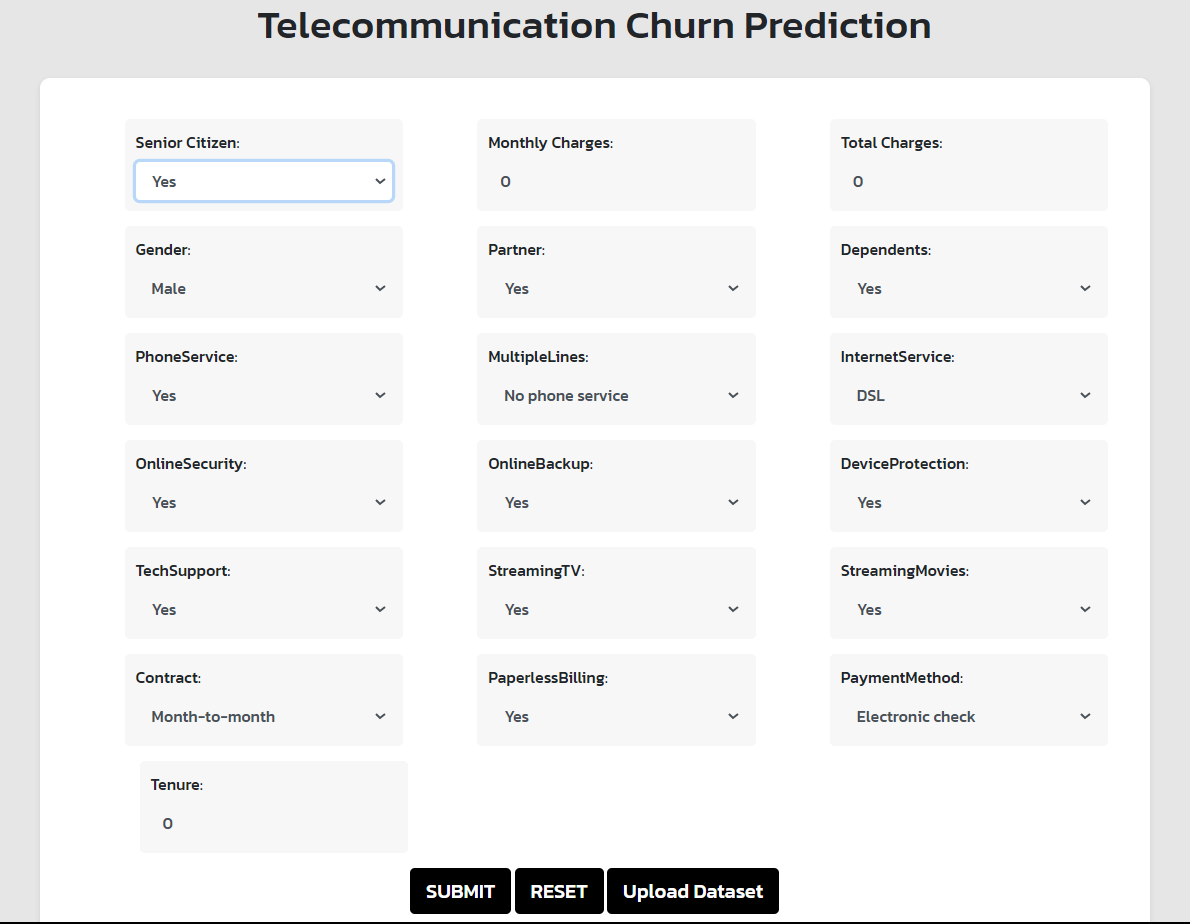
Description automatically generated**

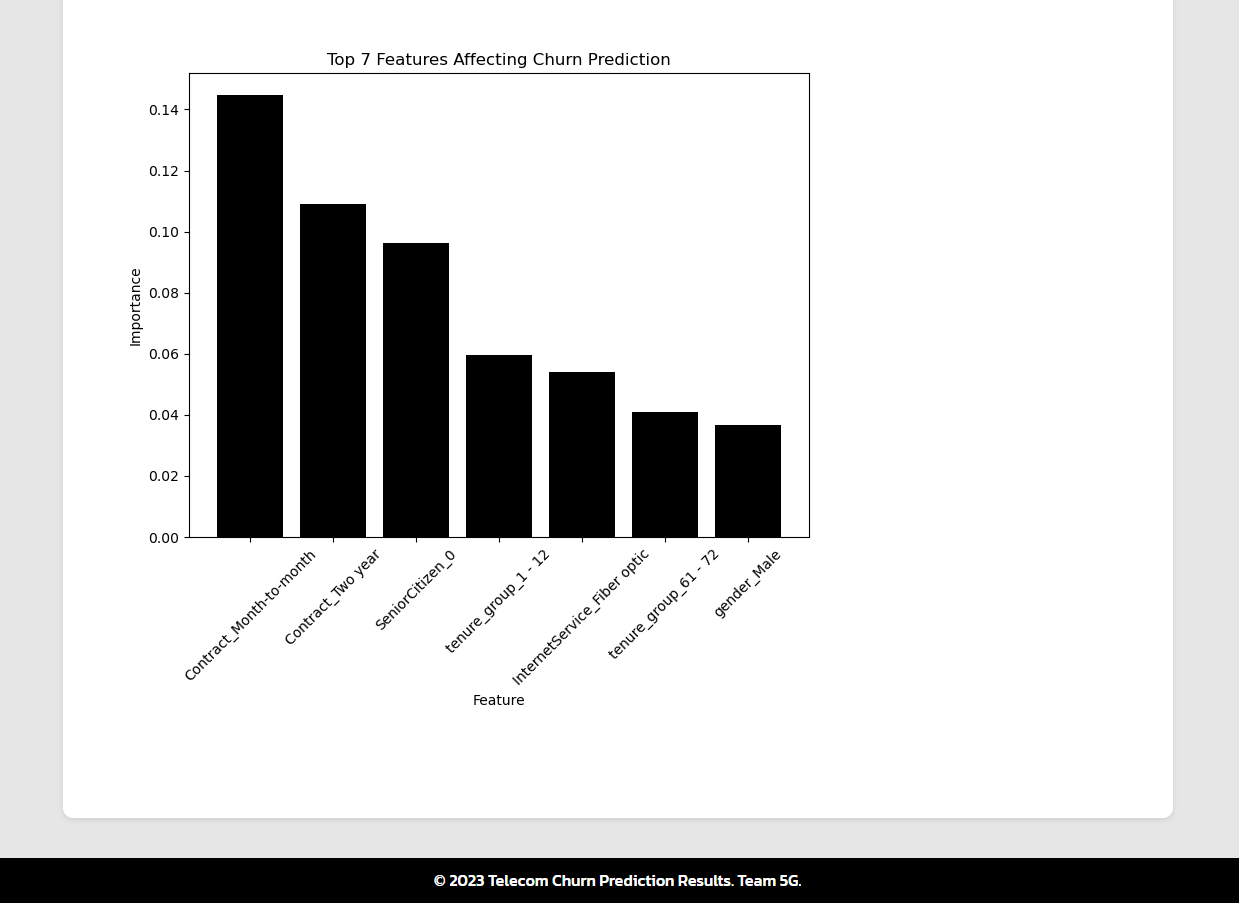
**OBSERVATIONS AND FINDINGS**

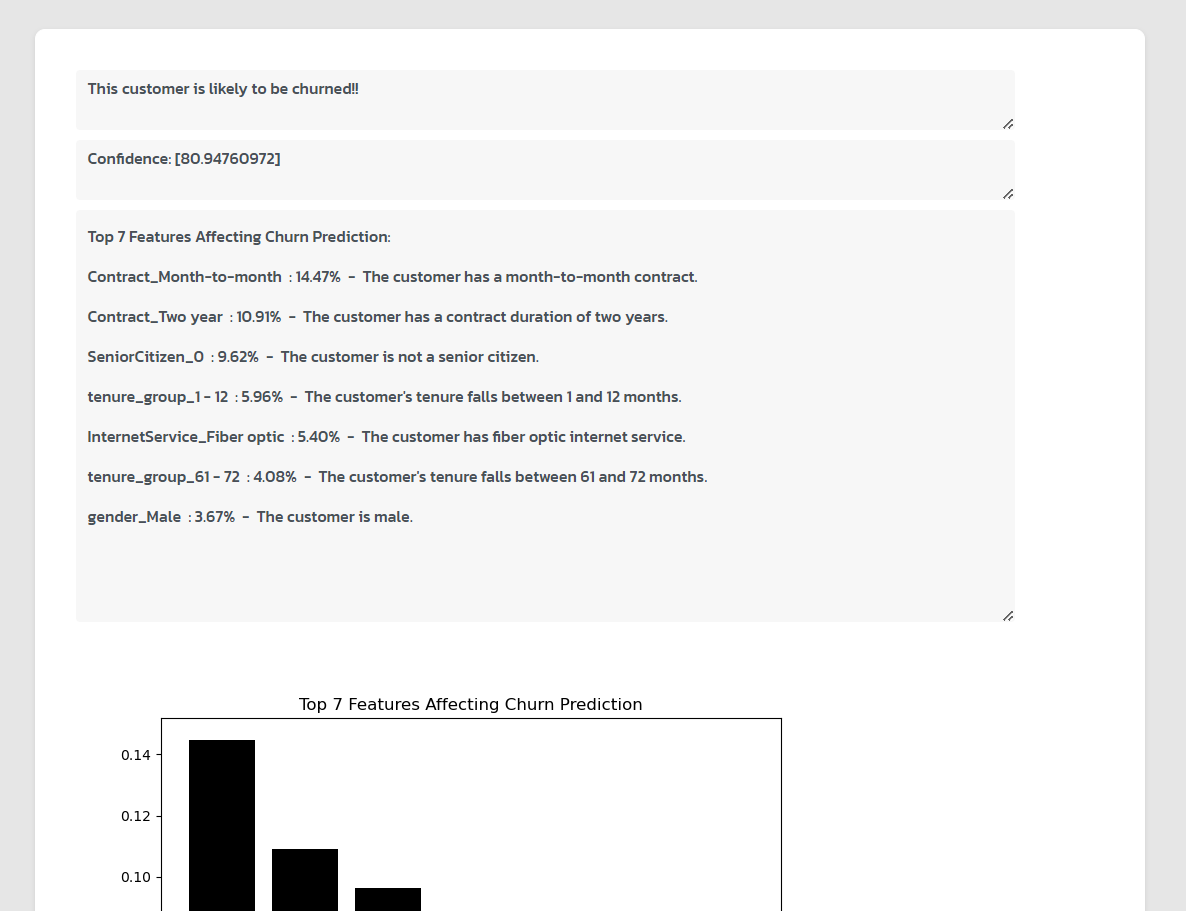
* + HIGH Churn seen in case of Month-to-month contracts, No online security, No Tech support, First year of subscription and Fiber Optics Internet
  + LOW Churn is seen in case of long term contracts, Subscriptions without internet service and the customers engaged for 5+ years.
  + Random Forest (after SMOTE) has the highest accuracy of 83%.

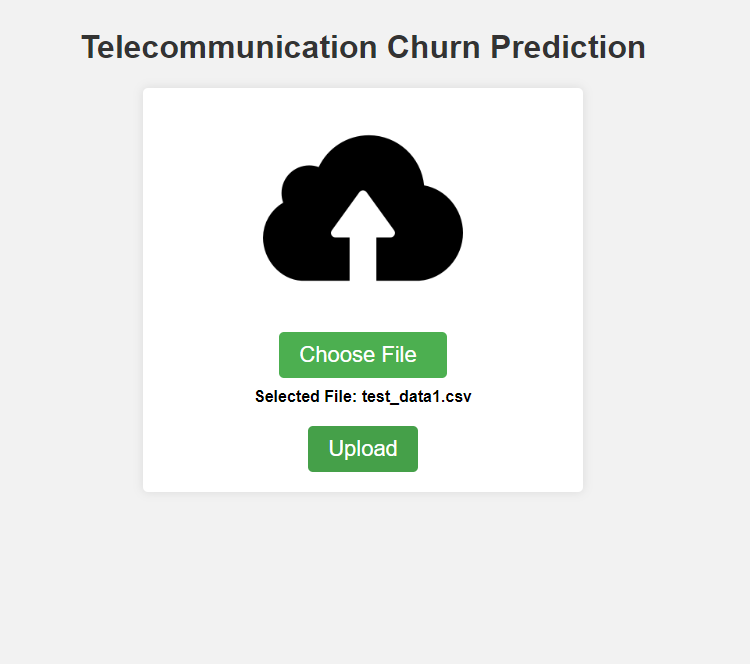
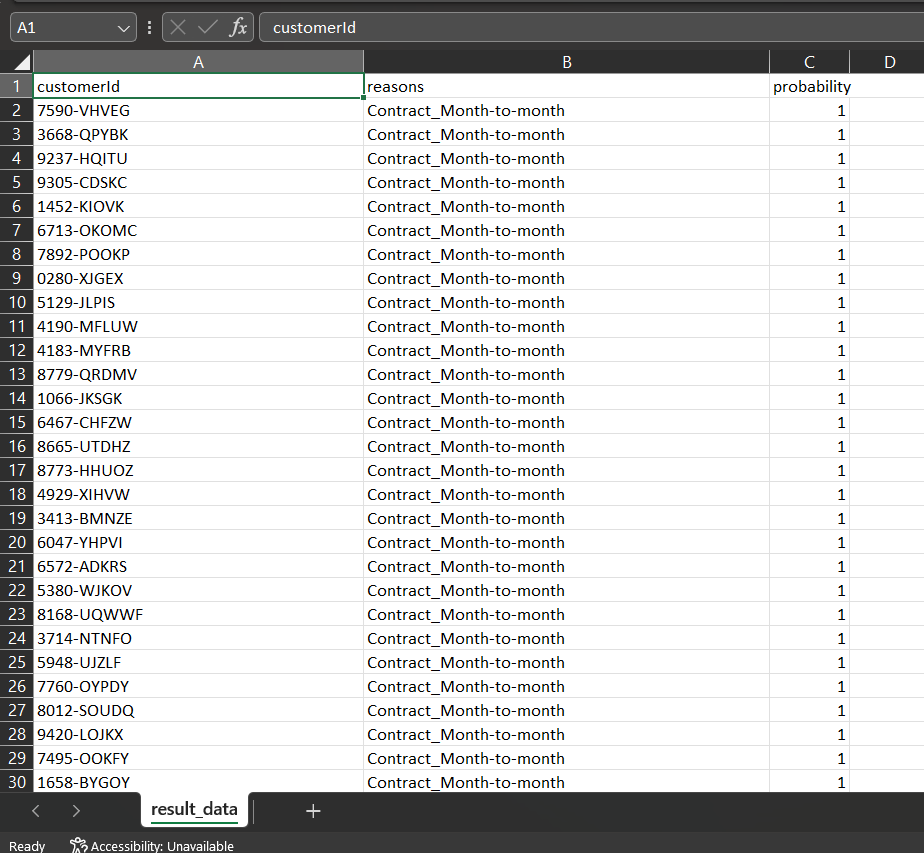
**7. RESULTS**

**7.1 OUTPUT SCREENS**









**8.TESTING**

**8.1. Types of Testing**

There are several types of testing that were performed on the "Telecom churn analysis" model to ensure its functionality and quality:

* Train-Test Split: This is the most basic form of testing, where the dataset is divided into training and testing sets to assess a model's generalization ability.
* Cross-Validation: K-fold cross-validation is widely used to obtain a robust estimate of a model's performance by splitting the data into multiple subsets and training/testing the model iteratively.
* Hold-Out Validation: A validation set is held out in addition to the training and testing sets to assess model performance during development.
* Feature Testing: Assess the importance of individual features using techniques like feature importance scores and permutation feature importance.
* Metrics for Evaluation: Common metrics such as accuracy, precision, recall, and F1 score are used to evaluate model performance based on the specific problem type.
* Model Fairness and Bias Testing: Evaluate the fairness and bias of the model using fairness metrics and bias mitigation techniques.

**8.2 Test cases:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test case id | Test title | Test data | Testing steps | Expected output | Actual output | status |
| TC001 | Predict Churn Probability | Input customer data for prediction | 1. Send a POST request to the Flask API with customer data. 2. Receive the prediction result. | Churn probability (float) | Churn probability (float) | Pass |
| TC002 | Detect High Churn Rate | Test dataset with known churn labels | 1. Send a GET request to retrieve churn rate from the Flask API. | Churn rate (float) | Churn rate (float) | Pass |
| TC003 | Invalid Input Data | Invalid customer data | 1. Send a POST request to the Flask API with invalid customer data. 2. Handle any error response. | Error message | Error message | Pass |
| TC004 | Model Re-training | New training data and parameters | 1. Trigger a re-training process for the model via a Flask API endpoint. | Success message | Success message | Pass |
| TC005 | Model Version Management | Multiple model versions | 1. Send a request to switch between different model versions via the Flask API. | Success message | Success message | Pass |
| TC006 | Unauthorized Access | Invalid or missing authentication | 1. Send a request without proper authentication headers. | Unauthorized error message | Unauthorized error message | Pass |

**9. Conclusion**

In summary, our project revolved around addressing a critical challenge in the telecom industry – customer churn. Telecom companies face the constant challenge of retaining their customers in a highly competitive market. To tackle this issue, we embarked on the journey of developing an effective machine learning model. Our model was designed to analyze extensive customer data, delving deep into various facets of user behavior and demographics. By scrutinizing this data, our model was able to identify the key factors contributing to customer churn. These factors encompassed everything from service usage patterns, customer complaints, billing information, and more. One of the standout achievements of our project was the impressive accuracy our model delivered. Through rigorous training and fine-tuning, we ensured that the model's predictions were not only reliable but also actionable. This meant that telecom companies could rely on the insights provided by our model to make strategic decisions and interventions.

The implications of our work were profound. Telecom companies now had a powerful tool at their disposal. They could proactively identify customers at risk of churning and take steps to retain them. This might include offering targeted promotions, improving customer service, or addressing specific pain points that the model highlighted. By implementing our model, telecom companies could enhance their customer retention strategies. This had a ripple effect, not only reducing customer attrition but also enhancing overall customer satisfaction. Satisfied customers are not only more likely to stay but also more likely to advocate for the company, potentially driving business growth through positive word-of-mouth. In essence, our project was a game-changer for the telecom industry. It empowered companies with the means to transform churn from a challenge into an opportunity. By using our machine learning model to identify and address the factors leading to churn, telecom companies could build stronger, more loyal customer bases, ultimately fostering growth and success in a highly competitive market.

**Future Scope**

1. **Real-time Churn Prediction:** Developing and implementing real-time churn prediction systems to enable immediate response and intervention when customers are at risk of churning.
2. **Advanced Machine Learning Techniques**: Exploring and incorporating advanced machine learning techniques, such as deep learning and ensemble methods, to enhance the accuracy and effectiveness of churn prediction models.
3. **Customer Segmentation:** Expanding and refining customer segmentation to tailor retention strategies to specific customer clusters, addressing their unique needs and preferences.
4. **Explainable AI:** Ensuring that churn prediction models are not only accurate but also explainable, providing insights into why the model makes specific predictions, which can aid in more targeted actions.
5. **User Interface and Visualization:** Developing user-friendly interfaces and visualization tools that enable telecom company staff to easily interpret churn predictions and take informed actions.

**References**

[1] Jesmi Latheef and S Vineetha, "LSTM Model to Predict Customer Churn in Banking Sector with SMOTE Data Preprocessing," in 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), 2021.

[2] Aisyah Larasati, Darin Ramadhanti, Yuh Wen Chen, and Abdul Muid, "Optimizing Deep Learning ANN Model to Predict Customer Churn," in 2021 7th International Conference on Electrical, Electronics and Information Engineering (ICEEIE), 2021.

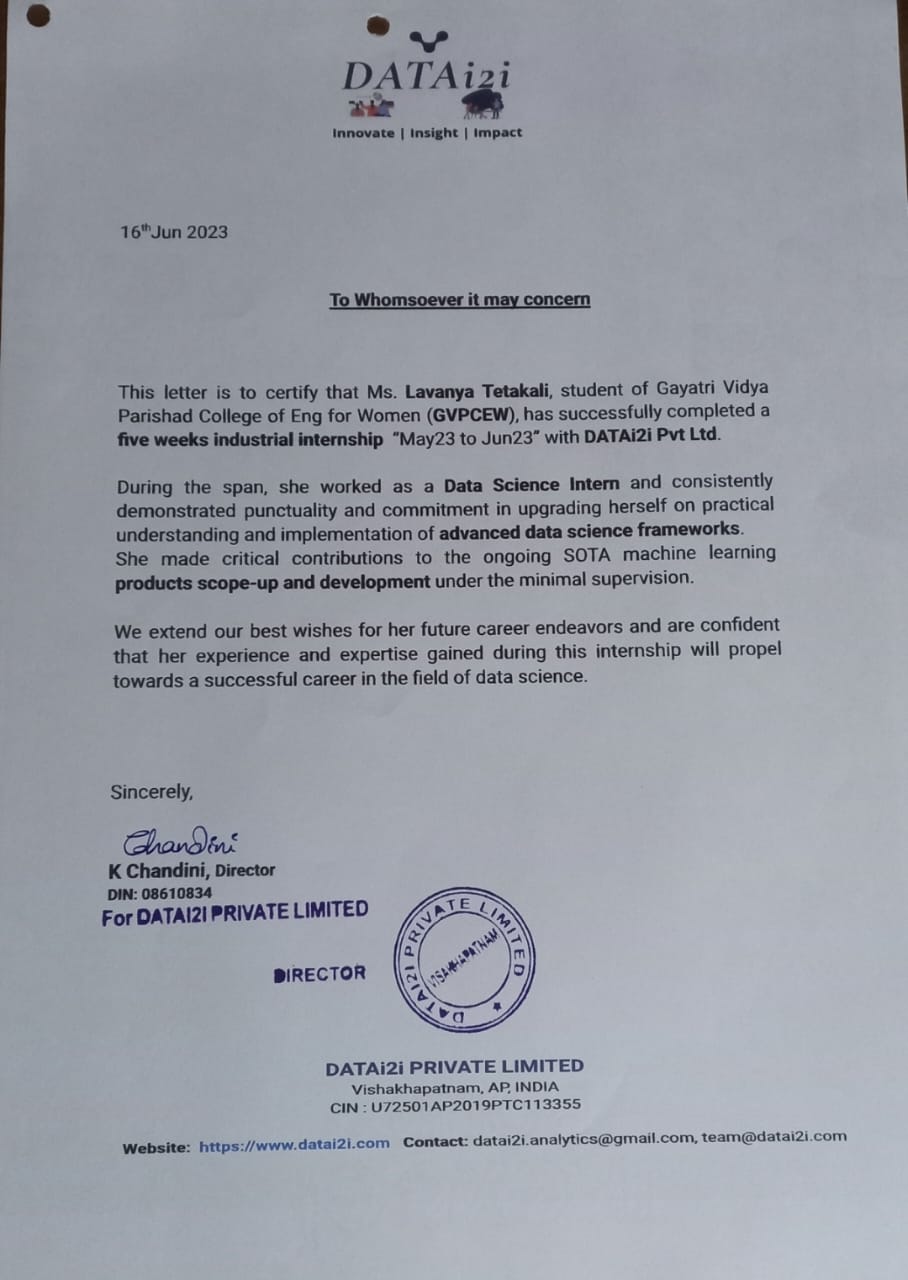
[3] Xiaobin Zhang, Gao Feng, and Huang Hui, "Customer-Churn Research Based on Customer Segmentation," in 2009 International Conference on Electronic Commerce and Business Intelligence, 2009.

[4] Peng Li, Siben Li, Tingting Bi, and Yang Liu, "Telecom customer churn prediction method based on cluster stratified sampling logistic regression," in International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014, 2014.

[5] Carl Gold, "Fighting Churn with Data: The Science and Strategy of Customer Retention," Manning, 2020.

[6] Bora Bardük, "Modelling Time Statistics for Customer Churn Prediction," in 2020 28th Signal Processing and Communications Applications Conference (SIU), 2020.

**Internship Certificate**

****