SMARTBRIDGE APPLIED DATA SCIENCE EXTERNSHIP

PROJECT REPORT

TITLE: "CUSTOMER CHURN PREDICTION"

TEAM MEMBERS:

1. **Name:** Piyush Mehar
   Reg No: 20BCE10065
   Campus: VIT Bhopal
   Email id: piyush.mehar2020@vitbhopal.ac.in

2. **Name:** Anubhav Shukla
   Reg No: 20BCE0708
   Campus: VIT Bhopal
   Email id: anubhavshukla.2020@vitbhopal.ac.in

1. INTRODUCTION:

1.1. Overview:

Retaining consumers is crucial to any business' success in today's fiercely competitive business environment. For businesses in all sectors, customer churn—the phenomenon of customers ending their engagement with a company—poses a serious challenge. Businesses can take preventative action and reduce customer attrition by foreseeing future churners. Machine learning is useful in this situation.

1.2. Purpose:

The purpose of this project is to use machine learning methods to build a precise and dependable customer churn prediction model. We can find hidden insights that cause churn by examining past customer data, including demographics, purchasing behavior, and interaction patterns. To create predictive models that may foretell client turnover, machine learning algorithms including, DecisionTree, Random Forests, Gradient Boosting, XGBoost, and Light Gradient Boosting Machine will be used.

This project offers two advantages. In order to execute targeted retention strategies and lower customer attrition, firms can first have a deeper understanding of the factors that influence customer churn. Second, by offering real-time churn risk scores and alerts through integration with the already-existing customer management systems, the predictive model enables businesses to act quickly to retain at-risk consumers.

By predicting customer churn with high accuracy, this project aims to equip businesses with actionable insights to enhance customer retention strategies, improve customer satisfaction, and ultimately drive long-term business growth.

2. LITERATURE SURVEY:

2.1. Existing Methods to Solve:

Existing methods or approaches to solve the problem of customer churn prediction typically involve traditional algorithms such as DecisionTree, Random Forests, Gradient Boosting, XGBoost, and Light Gradient Boosting Machine These methods often require interpretability, require handling non-linearity, and ensemble the methods for better prediction.

Decision Trees: Decision trees are another popular technique for churn prediction. They divide the data based on a series of hierarchical decisions to create a tree-like model. Each branch represents a decision based on a specific customer attribute, leading to a final prediction of churn or retention.

Random Forest: A machine learning algorithm used for customer churn prediction. It combines multiple decision trees to create an ensemble model, improving accuracy and handling large datasets. It analyzes customer data and predicts the likelihood of churn, helping businesses take proactive measures to retain customers.

Gradient Boosting: A machine learning algorithm for customer churn prediction. It iteratively builds an ensemble of weak decision trees, minimizing the loss function at each step. It effectively captures complex patterns, provides accurate predictions, and is widely used in industry for churn analysis and retention strategies.

XGBoost: A machine learning algorithm that can be used for customer churn prediction. It leverages gradient boosting techniques to create an ensemble of decision trees, enabling accurate identification of potential churners based on relevant features in a dataset.

LGBM: A powerful algorithm used for customer churn prediction. It efficiently handles large datasets, reduces memory usage, and provides high prediction accuracy. By leveraging gradient boosting, LightGBM can identify patterns and factors that contribute to customer churn, helping businesses take proactive measures.

Machine learning offers several advantages in customer churn prediction compared to traditional statistical or rule-based approaches. We can say this because with the help of machine learning scalability, adaptability of the dataset increases and enhances accuracy in various manners.

Machine learning algorithms are designed to learn patterns and relationships directly from data. They can uncover complex and non-linear patterns that may not be apparent through traditional analysis. This data-driven approach allows for a more accurate and comprehensive understanding of customer churn factors. Also, machine learning models can be continuously refined and improved as more data becomes available.

2.2. Proposed Solution:

The proposed solution for customer churn prediction using LGBM offers several advantages over traditional methods. With LGBM's approach, it leverages the strengths of ensemble learning to accurately identify customers at risk of churning.

Light Gradient Boosting Machine (LightGBM) algorithm is well-suited for customer churn prediction due to several reasons:

1. Efficiency: LightGBM is designed to be highly efficient, making it capable of handling large datasets with millions of records. It uses a histogram-based approach for binning, which reduces memory usage and speeds up training and prediction.

2. Accuracy: LightGBM leverages gradient boosting, a powerful ensemble learning technique, to iteratively build a predictive model by combining multiple weak learners. This allows it to capture complex patterns and interactions in the data, leading to high prediction accuracy.

3. Handling Imbalanced Data: Customer churn datasets often suffer from class imbalance, where the number of churned customers is significantly lower than non-churned ones. LightGBM provides techniques like weighted sampling and class balancing to effectively handle imbalanced data and prevent bias towards the majority class.

4. Feature Importance: LightGBM provides insights into feature importance, allowing businesses to identify the key factors that contribute to customer churn. This information helps in understanding the underlying drivers of churn and enables targeted retention strategies.

Overall, LightGBM offers a combination of efficiency, accuracy, and interpretability, making it a compelling choice for customer churn prediction tasks.

3. PROBLEM STATEMENT:

As a data scientist, The problem is to develop a machine learning model for customer churn prediction. The model will analyze customer data, including demographics, purchasing behavior, and engagement patterns, to accurately predict which customers are likely to churn. The objective is to provide businesses with actionable insights to implement targeted retention strategies, reduce customer attrition, and improve customer satisfaction, ultimately driving long-term business growth.

3.1 Proposed solution:

Our proposed solution for customer churn prediction utilizes a machine learning algorithm, specifically Light Gradient Boosting Machine. By leveraging the ensemble nature of Light Gradient Boosting Machine, we Preprocess and prepare the customer churn dataset, split the dataset into training and testing sets, train a LightGBM model on the training data, optimizing for churn prediction, evaluate the model's performance on the testing data using appropriate metrics (e.g., accuracy, precision, recall), fine-tune the model parameters to improve performance if necessary, use the trained LightGBM model to make churn predictions for new, unseen customer data, monitor and analyze churn predictions to identify potential churners and take proactive measures to retain them

To implement this solution, the following steps can be taken:

Downloading dataset: This is the first and the foremost data that needs to be done. The data has been downloaded from Kaggle and imported in the system within in the assigned folder.

Data Preparation: Clean the CSV (comma separated values) data by removing any duplicates, missing values, or inconsistencies. Ensure that the data are in a standardized format and the data in every row and column is understandable for us and the machine.

Data Preprocessing: It includes the import of required libraries, reading and analyzing the dataset, dropping unnecessary columns, changing column names if required, handling the missing values, encoding. After the data is sorted and is ready for the further process, splitting of data into dependent and independent variable is done. Lastly, we split the data into train and test data.

Training the model: With the help of the preprocessed data and by the usage of appropriate machine learning algorithm (in this case:), a model is built that will help us to achieve the solution to the problem statement.

Save the model: The model that was trained with the help of the preprocessed data need to be saved so that the solution can be achieved for the given problem statement like prediction, classification (in our case: prediction).

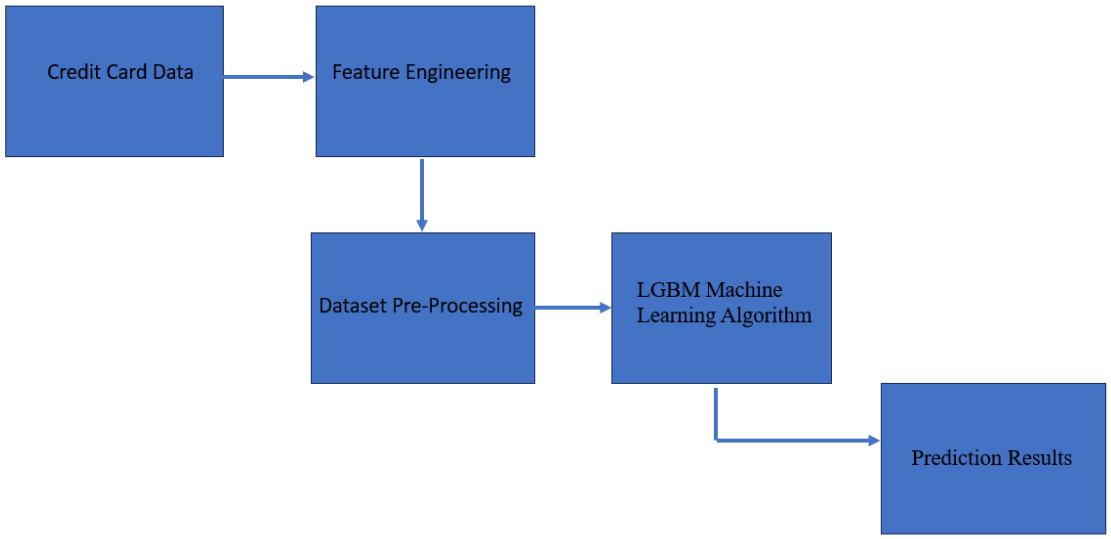
Web Application using Flask: Train and save your machine learning model using a suitable library, such as scikit-learn or TensorFlow. Set up a Flask application by creating a Flask app object and defining routes and views for handling HTTP requests. Load the saved model within the Flask

application, preferably during the app initialization phase. Define an endpoint for receiving input data via HTTP requests, such as a POST request. Extract the input data from the request and preprocess it if necessary. Pass the preprocessed data to the loaded machine learning model for prediction. Retrieve the prediction results from the model and format them appropriately. Return the prediction results as a response to the HTTP request.

By implementing this proposed solution, one can get to know about the customer's behavior that whether the customer will stay or leave.

4. THEORETICAL ANALYSIS:

4.1. Block Diagram:



4.2. Hardware and Software Requirements:

The hardware and software requirements for the customer churn prediction project using machine learning and Flask integration are as follows:

Hardware Requirements:

A computer or server capable of running machine learning algorithms, libraries and Flask smoothly, with sufficient processing power and memory.

Adequate storage capacity to store the dataset, Jupyter notebooks, and Flask web application files.

Reliable internet connectivity to access Jupyter notebooks and host the Flask web app.

Software Requirements:

Programming Language: Python is used in this case. Ensure you have a Python distribution installed, such as Anaconda or the official Python distribution from python.org.

Integrated Development Environment (IDE): An IDE provides an interactive coding environment and facilitates the development process. Popular options for Python include PyCharm, Spyder, Jupyter Notebook, and Visual Studio Code. Choose an IDE that suits your preferences and offers features like code editing, debugging, and project management.

Machine Learning Libraries: Install essential machine learning libraries such as NumPy, pandas, matplotlib, seaborn. These libraries provide various algorithms, tools, and utilities for data preprocessing, model training, and evaluation.

Data Manipulation and Analysis: Libraries like NumPy and Pandas are essential for data manipulation, cleaning, and exploratory data analysis. They provide efficient data structures and functions for working with numerical data and handling missing values.

Data Visualization: Matplotlib and Seaborn are popular libraries for data visualization in Python. It will allow us to create informative plots, charts, and graphs to gain insights from your data and communicate your findings effectively.

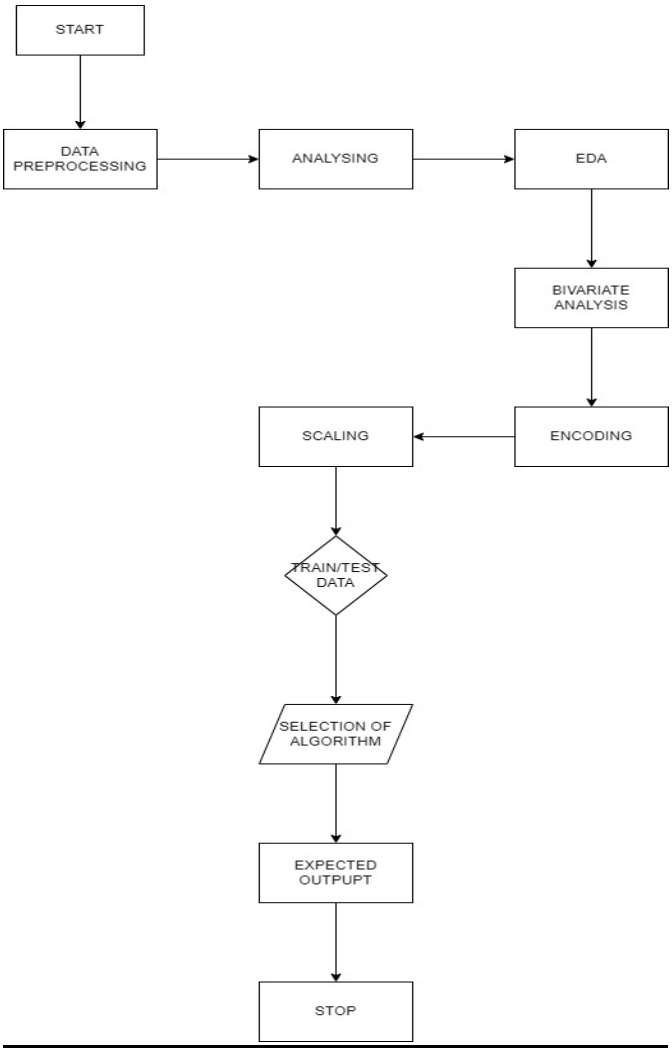
Web Framework: To integrate our churn prediction model into a web application, you might consider using a web framework like Flask. These frameworks enable us to build interactive web interfaces and handle HTTP requests for serving predictions to users.

Additionally, remember to install and manage these software requirements using package managers like pip or Conda to ensure version control and dependency management.

5. EXPERIMENTAL INVESTIGATIONS:

Experimental investigation in customer churn prediction using machine learning involves conducting empirical studies to evaluate and compare different machine learning algorithms, feature engineering techniques, and model evaluation metrics. It typically includes tasks such as data preprocessing, feature selection, model training and evaluation, parameter tuning, and performance analysis. The experiments aim to determine the most effective approach for accurately predicting customer churn, considering factors like prediction accuracy, model interpretability, computational efficiency, and scalability. Through systematic experimentation, insights are gained into the strengths and limitations of various machine learning techniques and their applicability to real-world customer churn prediction scenarios.

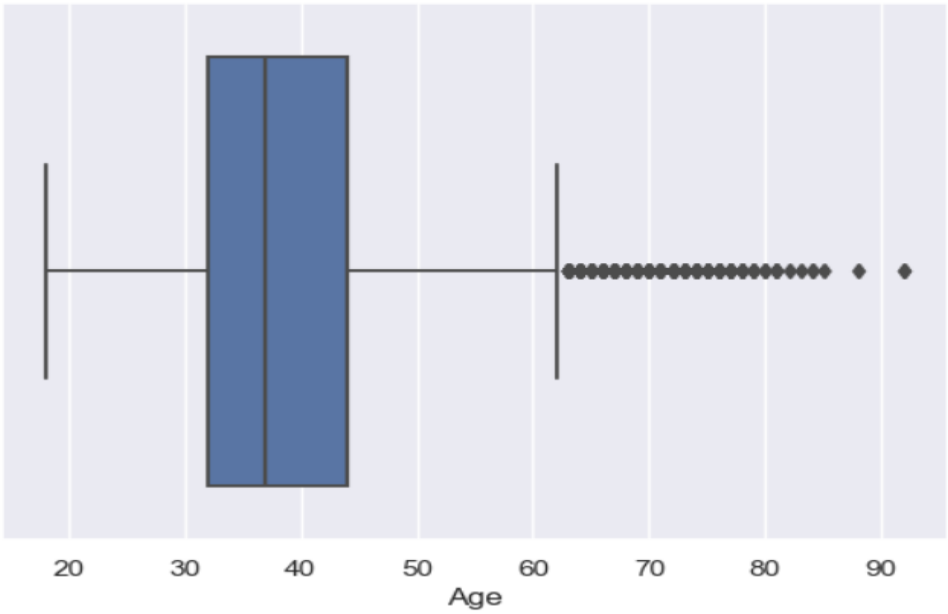
6. FLOWCHART:

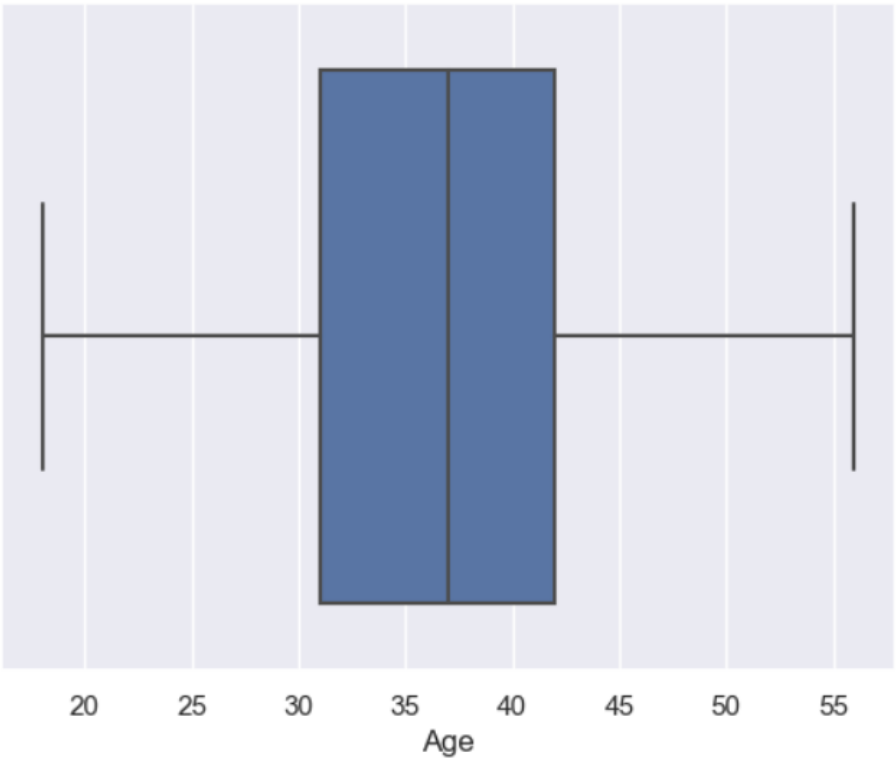


7. RESULT:

We have performed various data visualizations and have done bi-variate and multivariate analysis alongside the building of model. And also, we have integrated the model into a web application using Flask. Here are the results of our analysis:

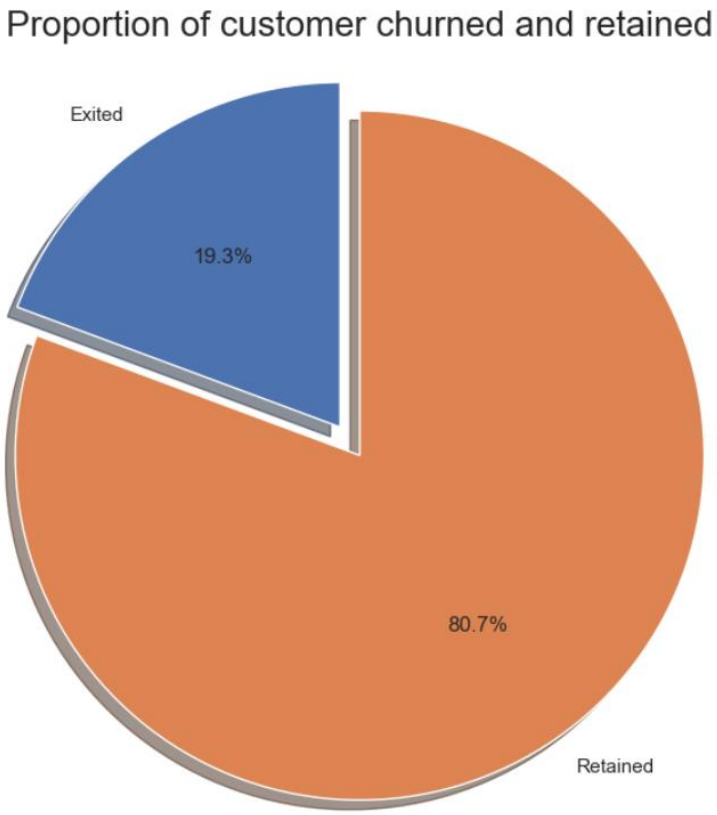
- Box Plot for age and removal of outliers



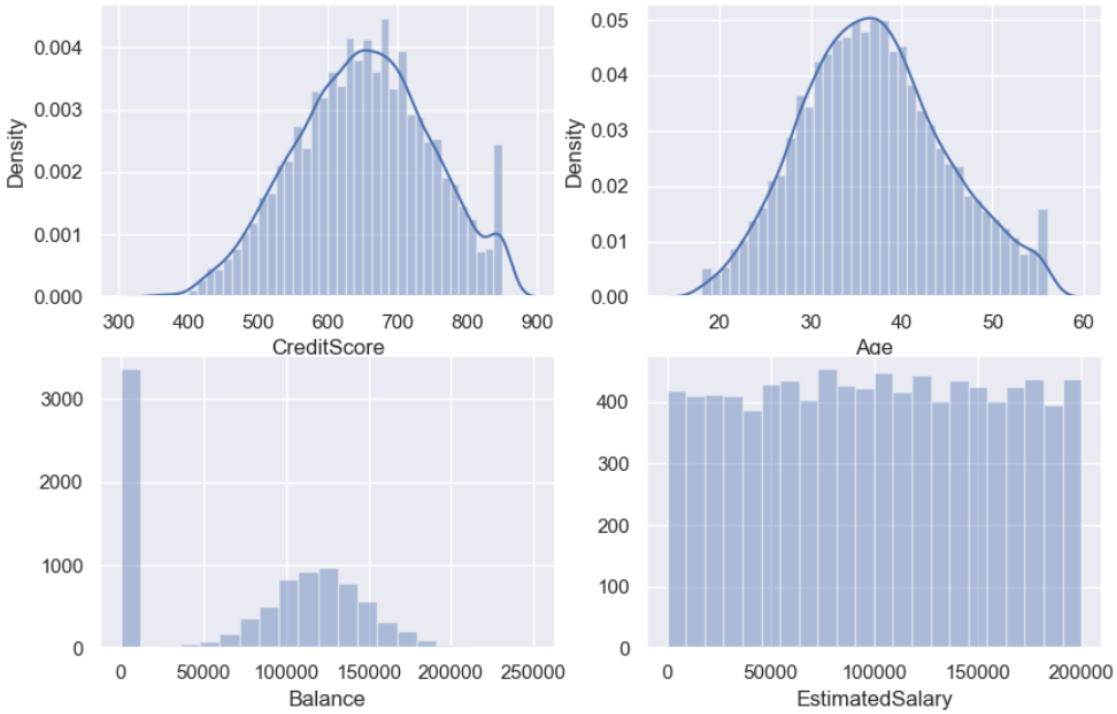


- Pie Chart to represent the number of people leaving or staying

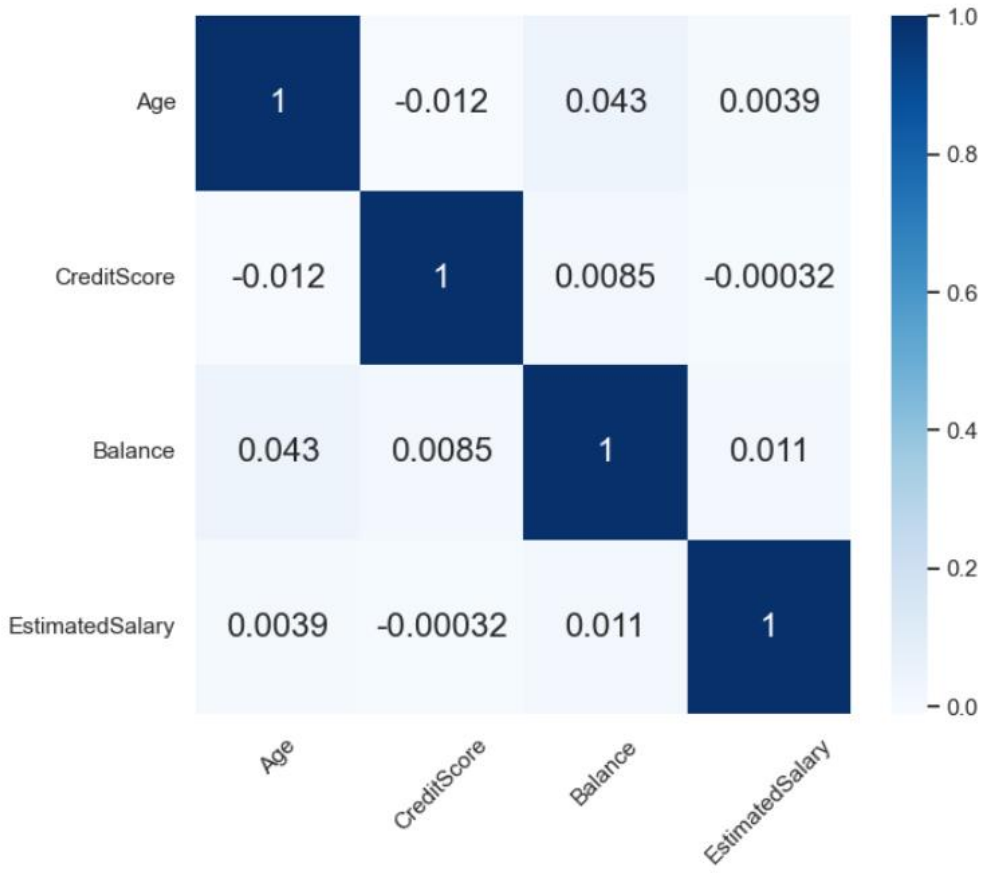
.



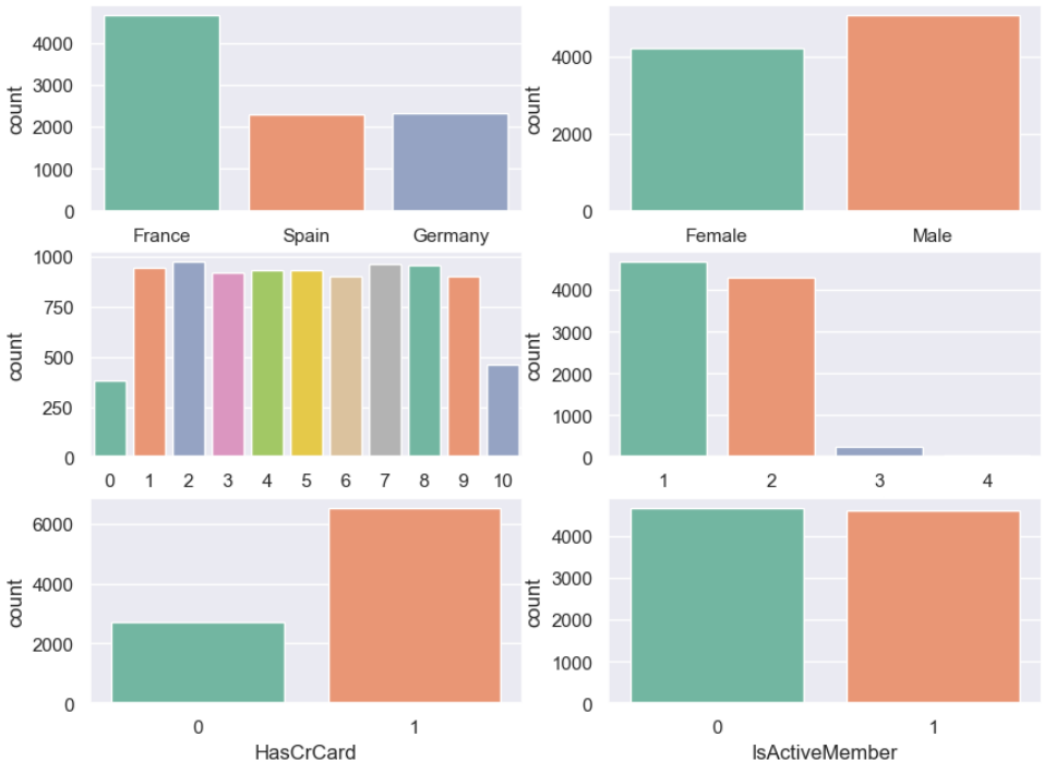
- Bar Plot to represent data of CreditScore, Age, Balance, EstimatedSalary



- Checking for correlation between Continuous Variables

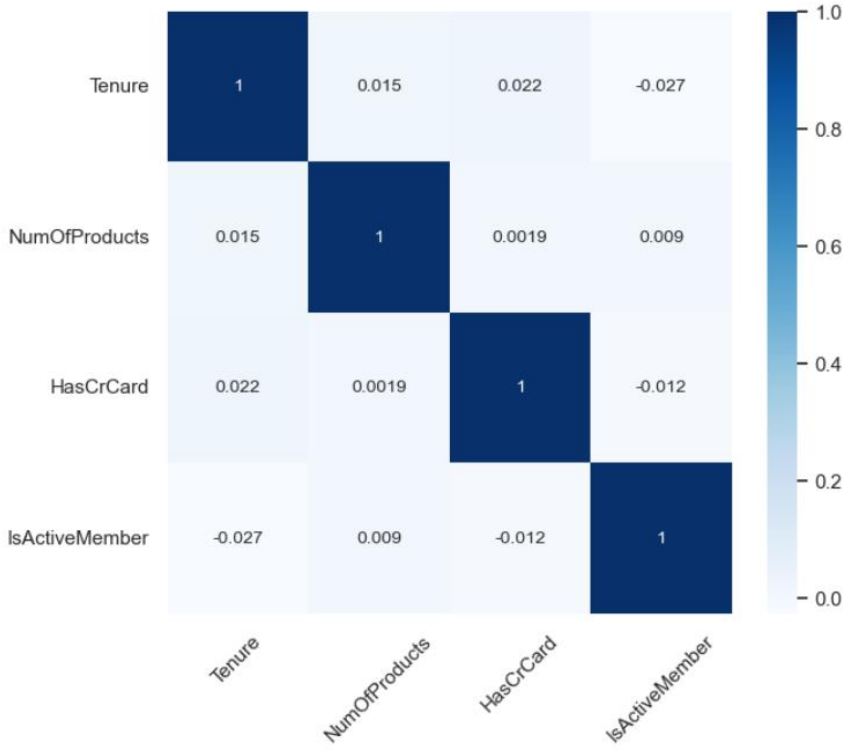


- Representation of Categorical Variables

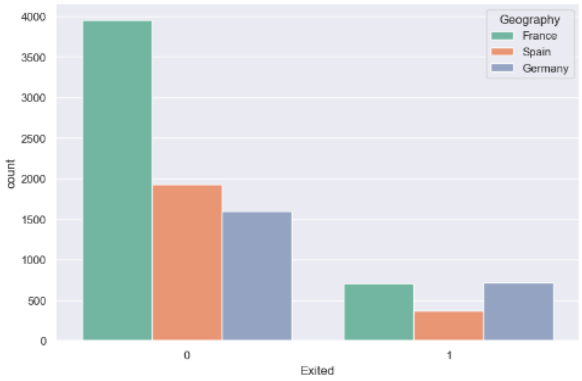


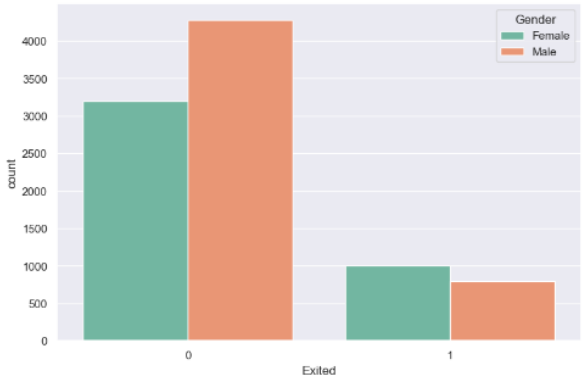
- Checking for corrleation between Categorical Variables

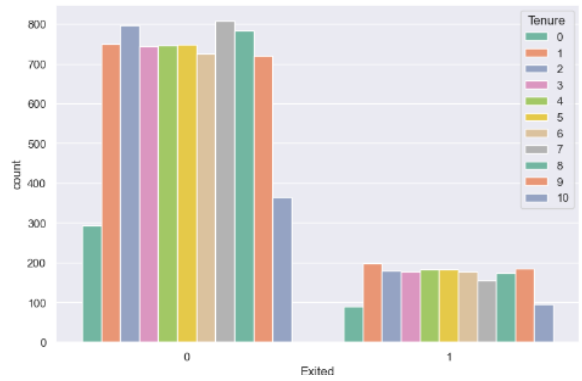
This chart depicts the trend of total assets over time, showing the changes in assets for each year.

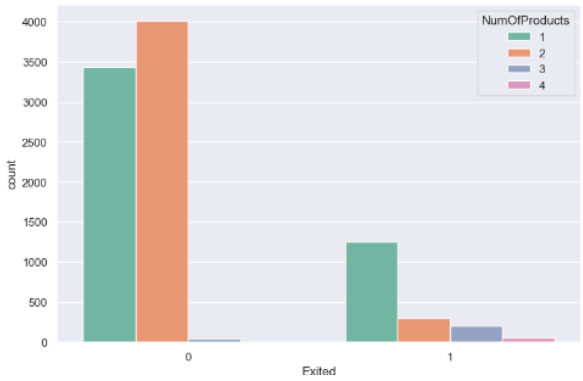


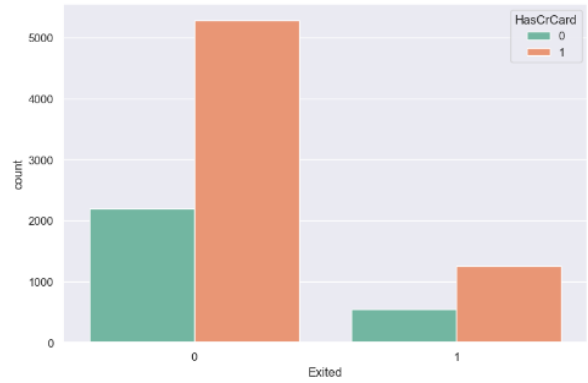
- Bivariate Analysis

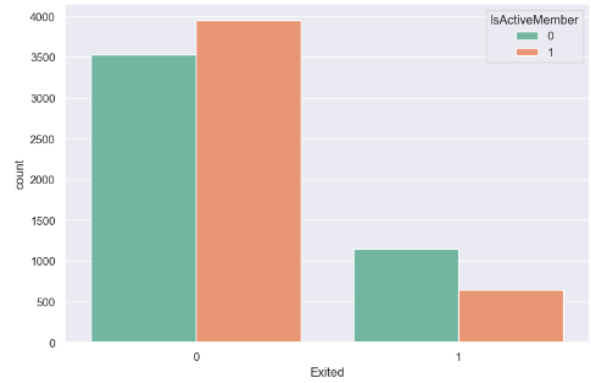




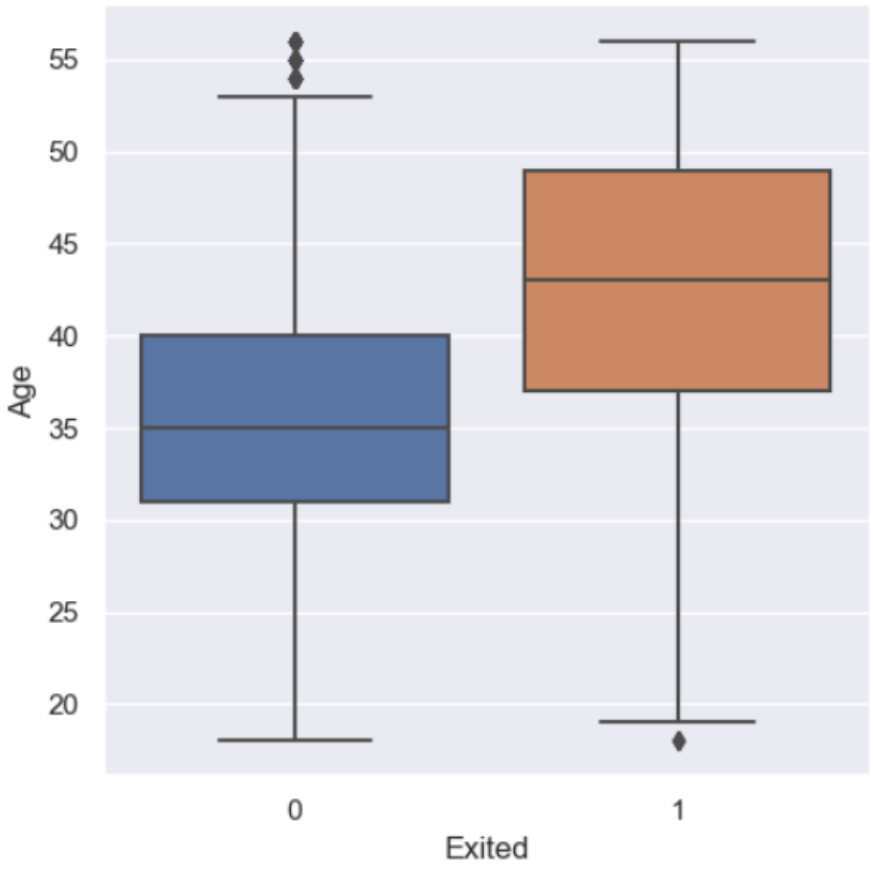








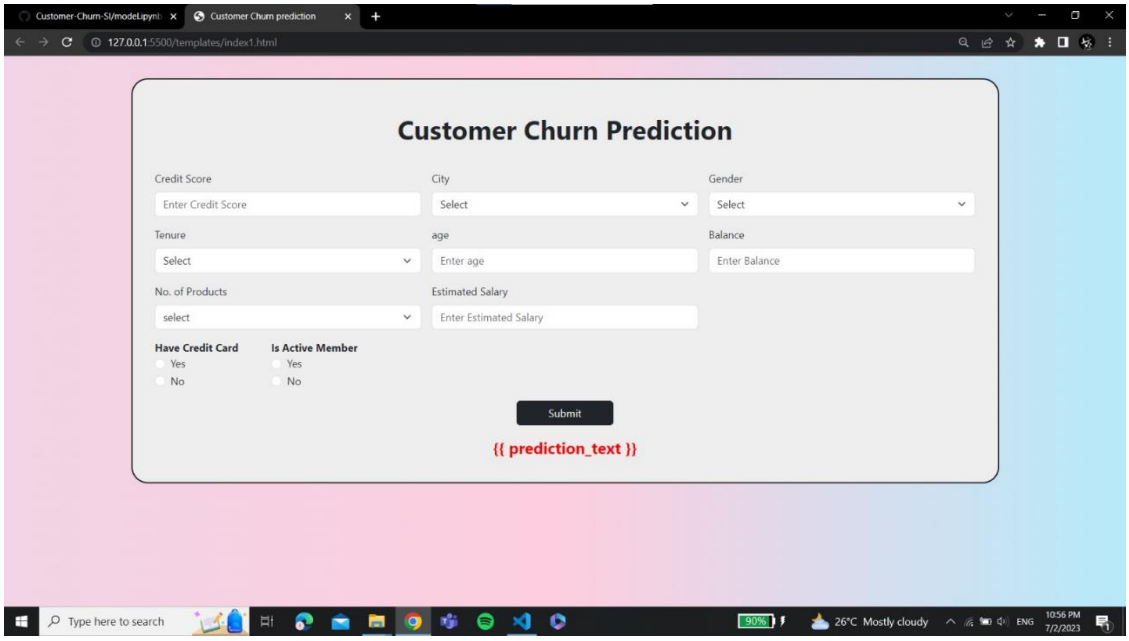
- Catplot to show the exited status of client

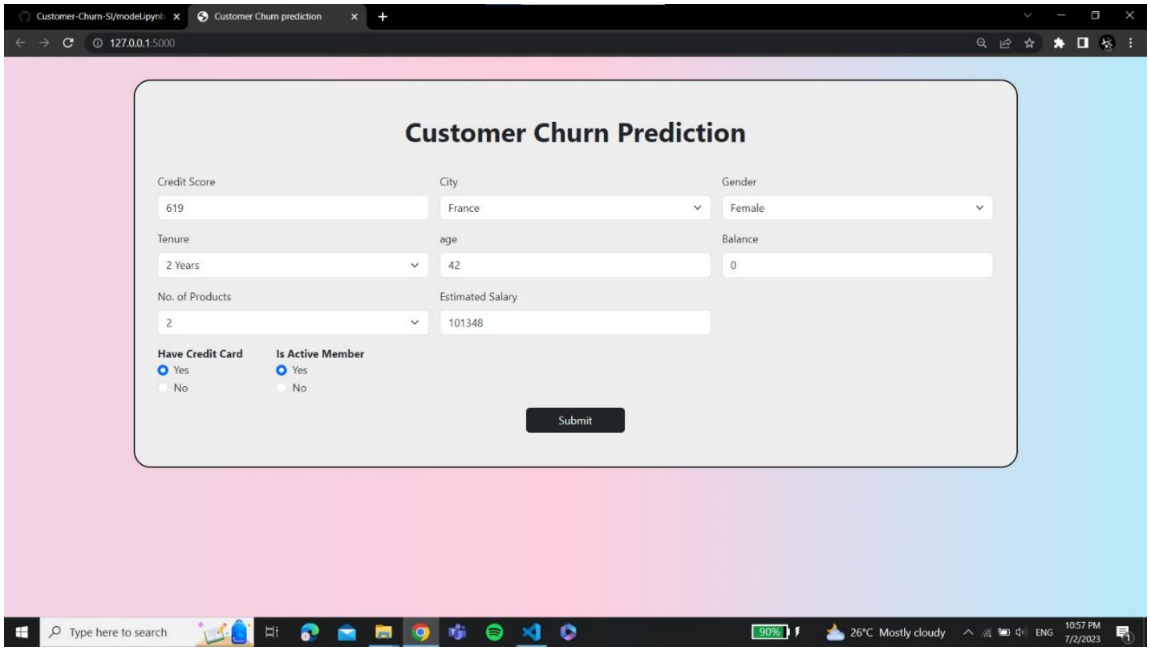


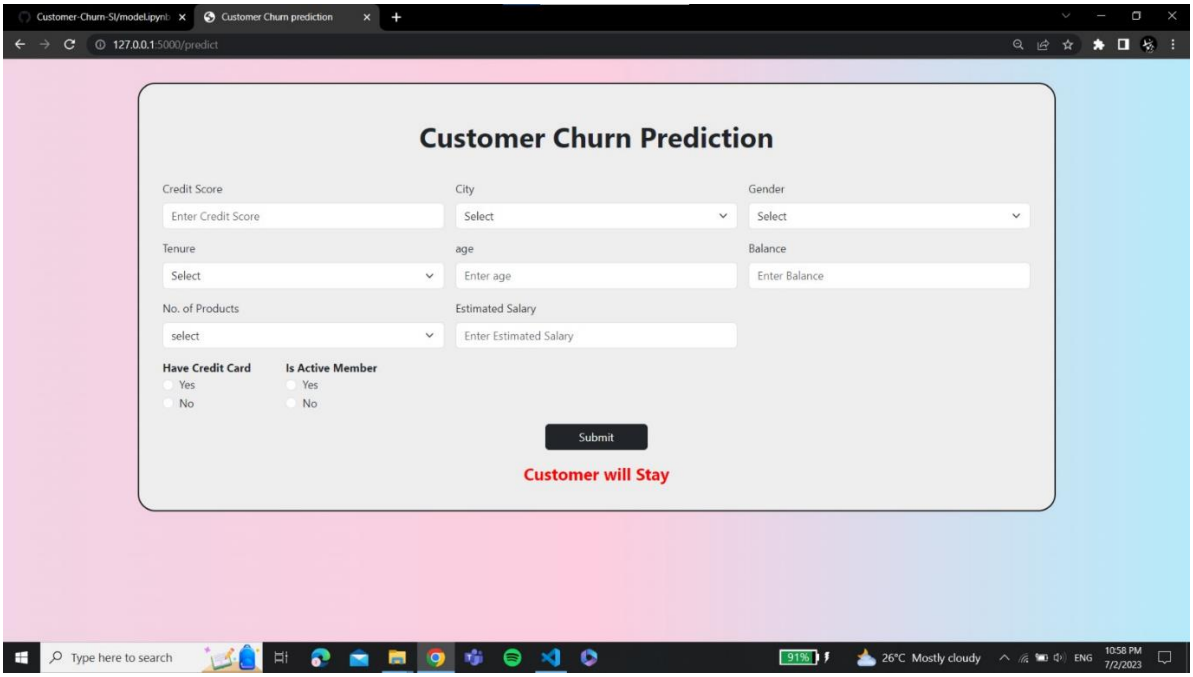
- Looking for Correlations

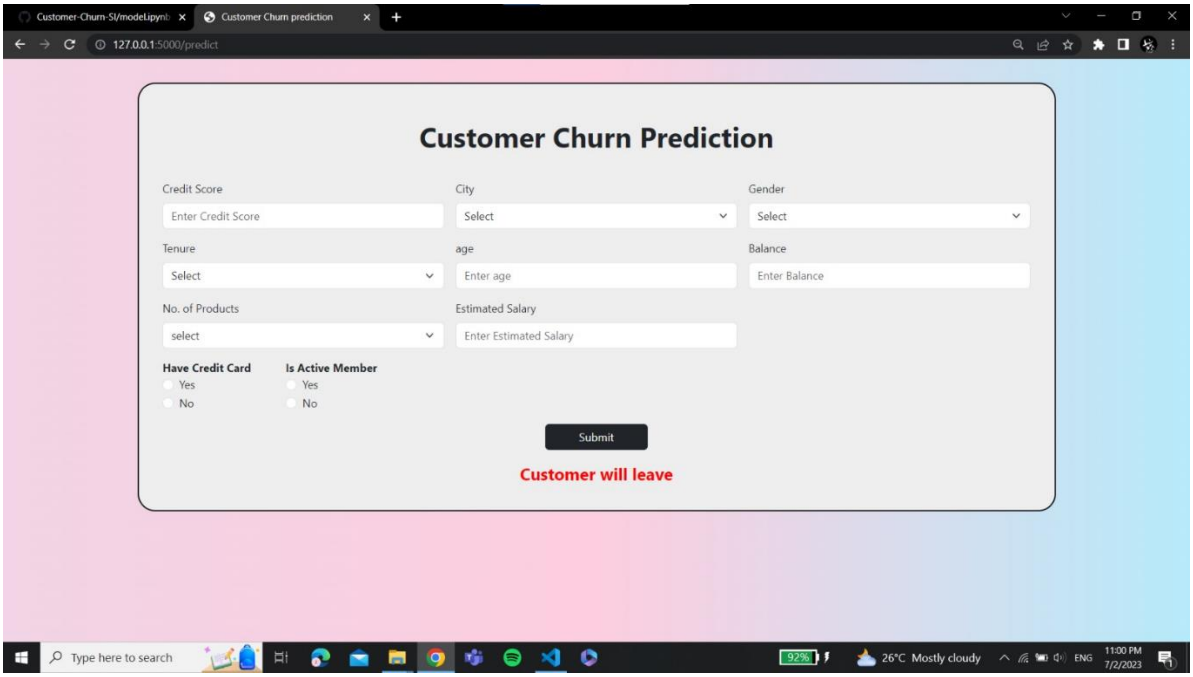
Furthermore, we have integrated the model into a web application using Flask, providing web-based access to the data and insights that we have generated.

Web Integration:






Customer-Churn-SI/modelipynb
Customer Churn prediction
127.0.0.1:5000/predict
Customer Churn Prediction
Credit Score
Enter Credit Score
City
Select
Gender
Select
Tenure
Select
age
Enter age
Balance
Enter Balance
No. of Products
select
Estimated Salary
Enter Estimated Salary
Have Credit Card
Yes
No
Is Active Member
Yes
No
Submit
Customer will Stay
Type here to search
91%
26°C Mostly cloudy
ENG
10:58 PM
7/2/2023


Customer-Churn-SI/modelipynb
Customer Churn prediction
127.0.0.1:5000/predict
Customer Churn Prediction
Credit Score
Enter Credit Score
City
Select
Gender
Select
Tenure
Select
age
Enter age
Balance
Enter Balance
No. of Products
select
Estimated Salary
Enter Estimated Salary
Have Credit Card
Yes
No
Is Active Member
Yes
No
Submit
Customer will leave
Type here to search
92%
26°C Mostly cloudy
ENG
11:00 PM
7/2/2023

8. VIDEO EXPLANATION

https://drive.google.com/file/d/1jtCmjD5nKlXAIUfvDoNjYX6mqkvdfBik/view?usp=sharing

9. ADVANTAGES & DISADVANTAGES:

Advantages:

- Improved Accuracy: Machine learning algorithms can uncover complex patterns and relationships in customer data, leading to more accurate churn predictions compared to traditional methods. This enables businesses to identify at-risk customers with higher precision.
- Actionable Insights: Machine learning models provide insights into the key factors driving customer churn. By understanding these factors, businesses can develop targeted retention strategies and take proactive measures to reduce churn rates.
- Scalability: Machine learning algorithms can handle large volumes of data and high-dimensional feature spaces, making them suitable for businesses with extensive customer bases and diverse sets of customer attributes.
- Real-time Predictions: Once trained, machine learning models can make churn predictions in real-time, enabling businesses to take immediate action and implement personalized retention strategies

- Automation: Machine learning models can be integrated into automated systems, allowing for continuous churn prediction and proactive retention efforts without manual intervention.

Disadvantages:

- Data Availability and Quality: Machine learning models rely on high-quality and representative data. Limited or poor-quality data can lead to inaccurate predictions and biased results.
- Interpretability: Some machine learning models, such as deep neural networks, are inherently complex and lack interpretability. It can be challenging to understand and explain the reasoning behind churn predictions, which may impact the trust and acceptance of the model's recommendations.
- Overfitting: Machine learning models are prone to overfitting if not properly regularized and validated. Overfitting occurs when the model performs well on the training data but fails to generalize to unseen data, leading to poor performance in real-world scenarios.
- Model Complexity and Computational Resources: Certain machine learning algorithms, particularly deep learning models, require significant computational resources and may have longer training times. This can pose challenges for businesses with limited computational capabilities.

- Implementation and Maintenance: Deploying and maintaining a machine learning-based churn prediction system requires technical expertise, infrastructure, and continuous monitoring and updating of models as new data becomes available.

10. APPLICATIONS:

The solution of using customer churn prediction using machine learning has various applications across different areas. Some of the key areas where this solution can be applied include:

❖ Telecom Industry
- Telecommunication companies can leverage churn prediction models to identify customers who are at a high risk of switching to a competitor.
- By identifying these customers in advance, companies can implement targeted retention strategies such as personalized offers or improved customer service to reduce churn rates.

❖ E-commerce:
- Online retailers can use churn prediction models to identify customers who are likely to stop purchasing from their platform.
- This information can be utilized to implement targeted marketing campaigns, offer personalized discounts, or improve the overall customer experience to retain those customers.

❖ Banking and Finance:
- Banks can employ churn prediction models to identify customers who are likely to close their accounts or switch to other financial institutions.
- By identifying these customers early on, banks can take proactive measures such as providing better interest rates, customized financial products, or improved customer service to retain them.

❖ Subscription-Based Services:
- Companies offering subscription-based services like streaming platforms, software-as-a-service (SaaS) providers, or online publications can use churn prediction models to identify customers who are likely to cancel their subscriptions.
- This information can be used to tailor retention strategies, such as offering personalized content recommendations, exclusive features, or discounted subscription plans.

❖ Insurance Industry:

- Insurance companies can utilize churn prediction models to identify policyholders who are at a high risk of not renewing their policies.
- By identifying these customers in advance, insurers can take measures such as adjusting premiums, providing additional coverage options, or offering personalized discounts to retain them.

❖ Health Care:

- Churn prediction models can be applied in the healthcare industry to identify patients who are likely to switch healthcare providers.
- This information can be used to implement proactive measures such as improving patient satisfaction, enhancing the quality of care, or offering personalized healthcare plans to retain patients.

11. CONCLUSION:

- ❖ Retention-focused strategies: Churn prediction models enable businesses to implement targeted retention strategies tailored to each customer. By understanding the factors that contribute to churn, companies can offer personalized incentives, discounts, or improved services to enhance customer satisfaction and loyalty.
- ❖ Cost savings: Acquiring new customers can be significantly more expensive than retaining existing ones. By identifying customers at risk of churn and implementing retention strategies, businesses can reduce the financial burden associated with acquiring new customers and maximize their return on investment.
- ❖ Improved customer satisfaction: By addressing the needs and concerns of at-risk customers before they churn, businesses can enhance customer satisfaction. Offering personalized solutions or addressing specific pain points can strengthen the customer-business relationship, leading to increased loyalty and positive word-of-mouth referrals.
- ❖ Business performance optimization: Churn prediction models provide valuable insights into customer behavior, preferences, and satisfaction levels. This information can be used to optimize product offerings, marketing campaigns, and customer service processes, ultimately improving overall business performance.

12. FUTURE SCOPE:

- ❖ Improved Accuracy: Machine learning algorithms will continue to evolve, resulting in improved accuracy in predicting customer churn. Researchers and practitioners are constantly working on developing advanced models that can effectively identify patterns and factors leading to churn, leading to more accurate predictions.
- ❖ Utilizing Unstructured Data: Currently, most churn prediction models rely on structured data like customer demographics and transaction history. However, there is a vast amount of unstructured data available from sources such as social media, customer reviews, and customer support interactions.
- ❖ Integration of Multiple Data Sources: Organizations are increasingly collecting data from various sources, including CRM systems, call center logs, website interactions, and more. Future models will integrate data from multiple sources to gain a holistic view of customer behavior and improve churn prediction accuracy.

- ❖ Integration with Business Processes: Churn prediction models will be seamlessly integrated into business processes, allowing organizations to take proactive actions based on predictions. Automated workflows will be created to trigger appropriate interventions, such as personalized offers or targeted marketing campaigns, to retain at-risk customers.

13. BIBILIOGRAPHY:

https://arxiv.org/ftp/arxiv/papers/1912/1912.11346.pdf

https://ieeexplore.ieee.org/document/8389557/figures#figures

https://www.semanticscholar.org/paper/Customer-churn-prediction-for-retail-business-Patil-Deepshika/e0057377a3c1b92b94854668e179ca06c3e39cc8

https://www.kaggle.com/code/a165079/credit-card-customer-churn-prediction