Engage Ai

CHURN PREDICTION & RETENTION SYSTEM

RAJU KUMAR

B.TECH (NIT DURGAPUR)

Table of Content

Problem Statement

The task is to design and implement a churn prediction and retention system for a company.

The solution should include:

- **Churn Prediction Models:** Analyze customer data to predict churn and identify contributing factors.
- **NLP for Customer Feedback Analysis:** Analyze textual feedback to understand customer dissatisfaction.
- **Generative AI for Retention:** Develop personalized retention strategies to reduce churn.

Customer Dataset - attached

Data Understanding & Preprocessing

- **Preprocessing steps**

Data understanding → Handle missing values → Remove Duplicates → Handle outliers → Encode categorical variables → Feature scaling & Engineering

- **Examine the provided datasets (Dataset Overview)**

Shape: 999 rows × 23 columns.

Data Types:

- Numerical: Senior Citizen, tenure in months, Monthly Average Balance (USD), Recommendation.
- Categorical: Gender, Marital Status, Dependents, Priority Account, Credit Cards, Loan Account, etc.

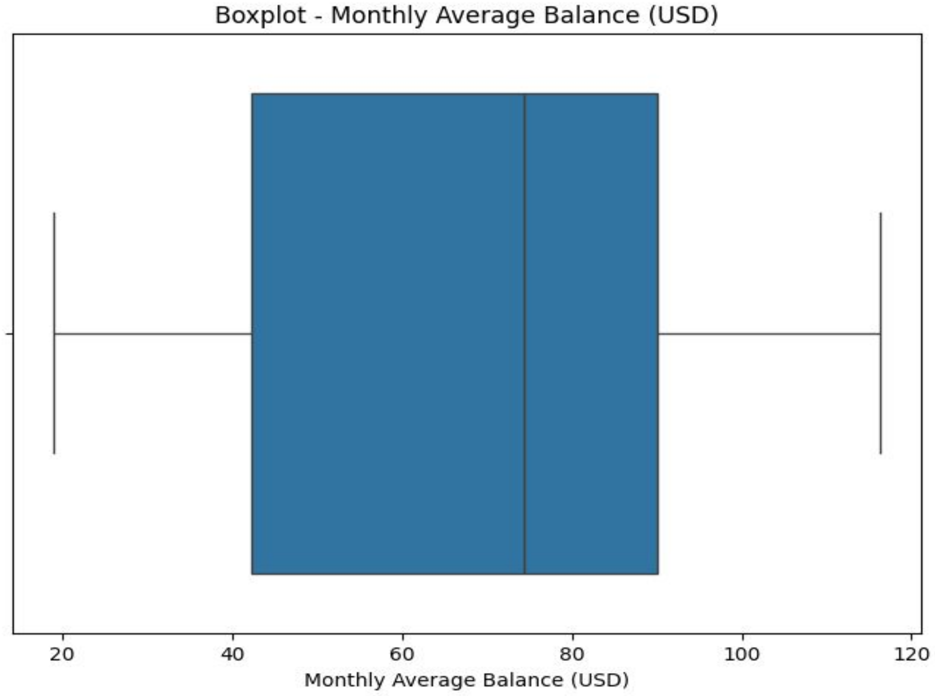
Target Variable: Churn (Categorical)

Handling Missing Values (Use Imputation)

Imputation Techniques: Use mean/median for numerical data, mode for categorical data, forward/backward fill for time series, or drop rows/columns with excessive missing data.

Handling Outliers

1. **Identifying Outliers:** Use box plots, z-scores, or IQR methods.
2. **Outlier Treatment:**
   - Cap/Clamp: Replace outliers with threshold values.
   - Transformation: Apply transformations (e.g., log, square root) to mitigate impact.
   - Exclusion: Remove outliers from the dataset when necessary



Data Transformation

1. **Encoding Categorical Variables**: Convert categorical data to numerical using techniques like Label Encoding
2. **Feature Scaling**: Normalize or standardize numerical features using StandardScaler to bring them to the same scale,
3. **Feature Engineering**: Create new features based on existing data to improve model performance.

Exploratory Data Analysis to Get Insights

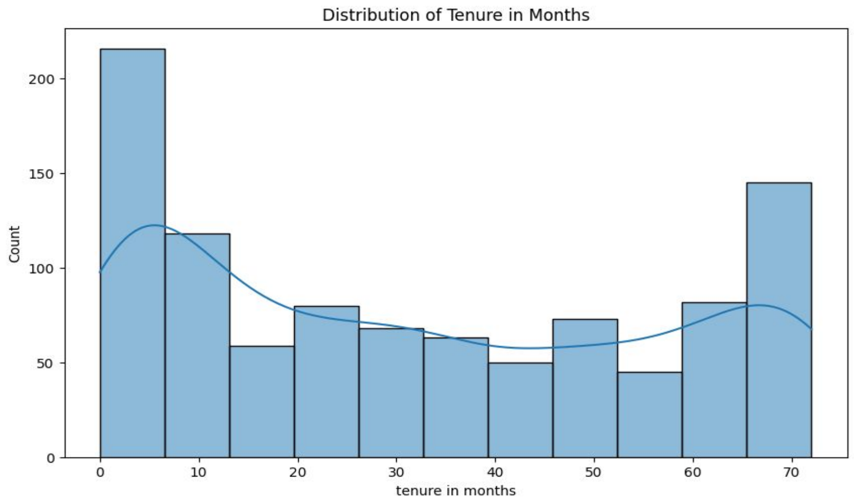
Insights-

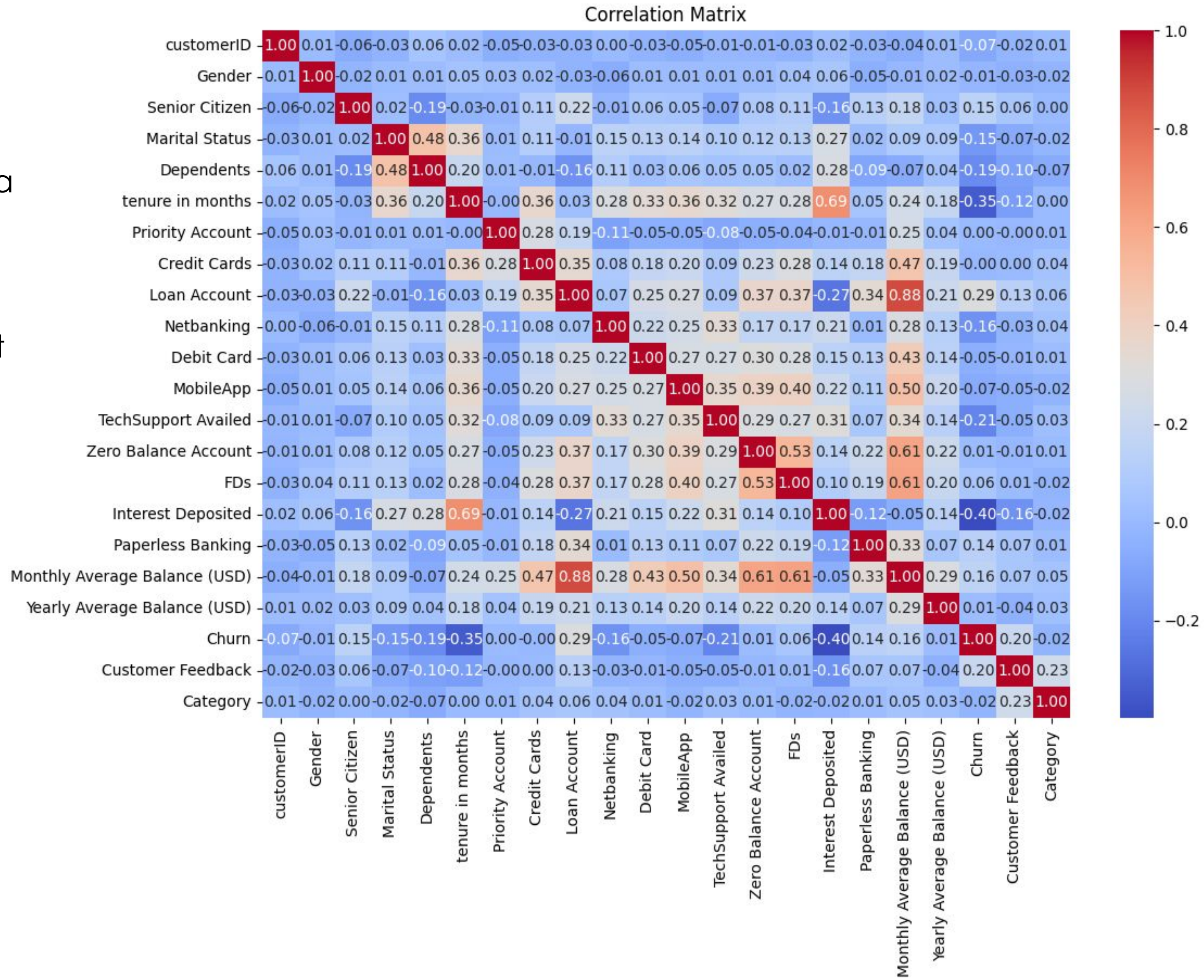
1. **High Correlation (Positive):**

   The ‘Loan Account’ feature has the strongest positive correlation with churn, suggesting it's a key factor in customers deciding to leave.

2. **Moderate Correlation (Negative):**

   Features like ‘Tenure in months’, ‘Tech Support Availed’, ’Interest Deposited’, and ‘Paperless Banking ’ are moderately negatively correlated with churn, indicating they may help in retaining customers.





Models for Churn Analysis

Logistic Regression Model-

Logistic Regression is a statistical method used for binary classification, estimating the probability that an input belongs to one of two classes, such as churn or no churn.

Logistic Function (Sigmoid): Logistic Regression uses the sigmoid function to map the predicted values (log-odds) to a probability between 0 and 1.

$\sigma(z) = \frac{1}{1+e^{-z}}$ where z is a linear combination of input features

(i.e., $z = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn$)

Applications: Commonly used in customer churn prediction, credit scoring, and medical diagnosis.

```
Logistic Regression Model Accuracy: 0.745

Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       146
           1       0.54      0.35      0.43        54

    accuracy                           0.74       200
   macro avg       0.67      0.62      0.63       200
weighted avg       0.72      0.74      0.73       200
```

Random Forest Model-

Random Forest is an ensemble learning method that builds multiple decision trees during training and merges their outputs to improve accuracy and control overfitting. It Handles non-linear relationships well.

Key Concepts:

- Decision Trees: Base models that split data using criteria like Gini impurity or entropy.
- Bagging (Bootstrap Aggregating): Multiple subsets of the data are created by random sampling with replacement.
- Random Feature Selection: At each split, a random subset of features is considered, reducing model variance.

```
Random Forest Model Accuracy: 0.8

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.92      0.87       146
           1       0.69      0.46      0.56        54

    accuracy                           0.80       200
   macro avg       0.76      0.69      0.71       200
weighted avg       0.79      0.80      0.79       200
```

Prediction Process:

- Majority Voting: Each tree votes for a class (e.g., churn/no churn), and the final prediction is based on majority vote.

Feature Importance:

- Random Forest provides insights into feature importance, showing which factors most influence predictions.

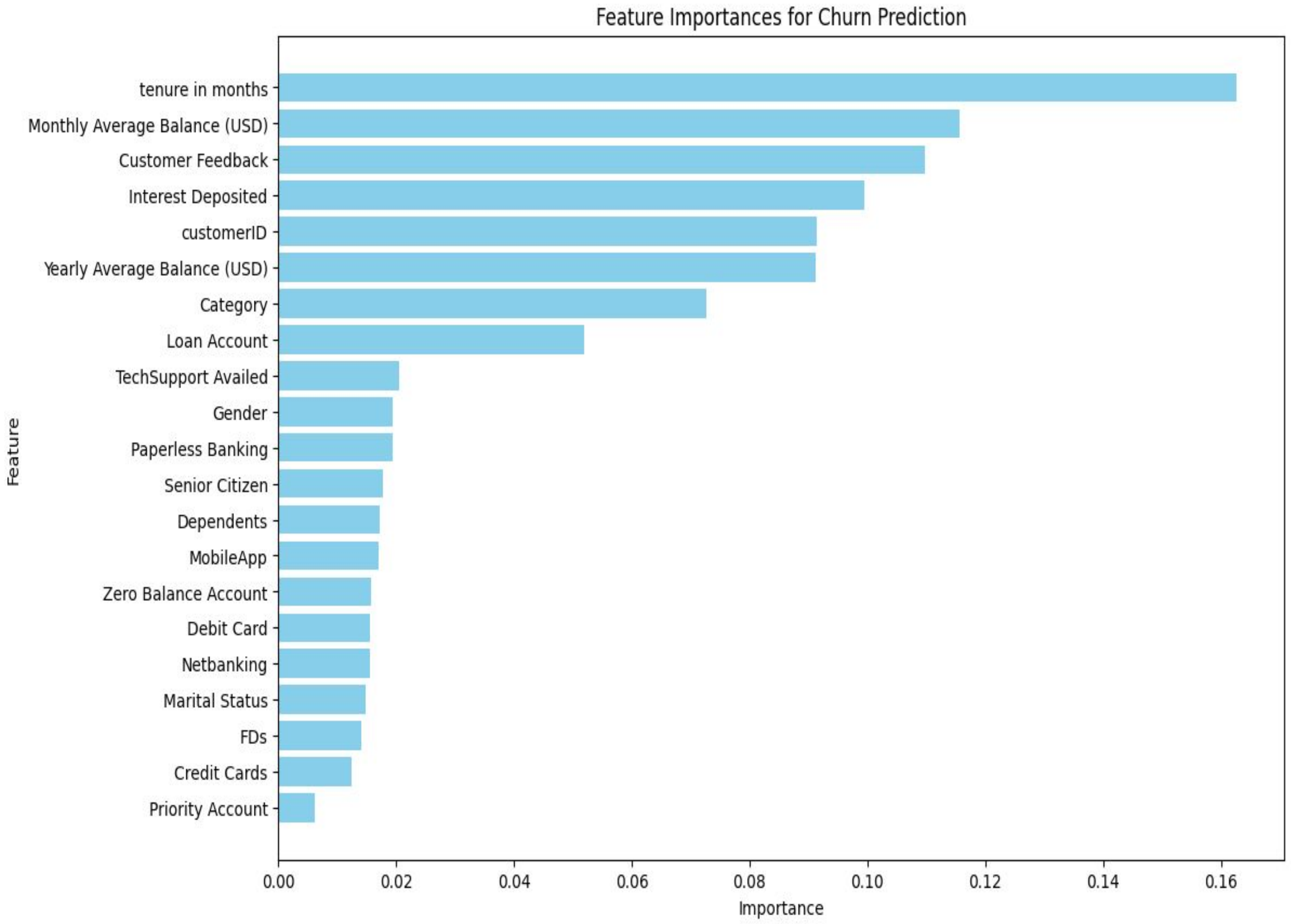
Feature Importance & Their Contribution to Churn

Importance by Random Forest:

- Measures feature importance by calculating the average decrease in impurity (e.g., Gini impurity) across all trees.

Example Features:

- Tenure in Months: Longer tenure may indicate loyalty.
- Monthly Average Balance: Higher balances could correlate with lower churn.
- Service Usage: Features like Netbanking or MobileApp usage can reflect customer engagement.



NLP for Customer Feedback Analysis

Sentiment Analysis to Gauge Overall Customer Satisfaction

```
Sentiment Summary:
count    999.000000
mean       0.148403
std        0.300404
min       -0.757900
25%        0.000000
50%        0.000000
75%        0.401900
max        0.875000
Name: Sentiment, dtype: float64
```

Objective: Assess customer sentiment from feedback.

Process:

1. Preprocessing: Clean and prepare feedback text.
2. Sentiment Analysis:
   - Library: **SentimentIntensityAnalyzer** from **VADER**.
   - Score Calculation: Uses VADER's **polarity_scores()** method to determine sentiment, providing a composite score ranging from -1 (most negative) to 1 (most positive).

Outcome:

- Sentiment Summary: Metrics include mean, standard deviation, and range of sentiment scores.
- Insight: Understand customer sentiment (positive, negative, or neutral) to identify trends and areas for improvement.

Extracting Reason for Dissatisfaction

Topic Modeling with Latent Dirichlet Allocation (LDA)

Purpose: Extract meaningful topics from customer feedback to understand common themes.

Process:

- <u>Identify Topics:</u> LDA assigns feedback to topics based on word distributions.
- <u>Display Top Words:</u> display_topics() function lists the most significant words for each topic.
- <u>Outcome:</u> Each topic is characterized by a set of prominent words, revealing underlying themes in customer feedback.

```
Topics Extracted:

Topic 1:
loan processing long process time opening bank account activation easy

Topic 2:
customer service current account deposit fixed helpful support cards credit

Topic 3:
card debit process credit loan slow pin confusing complicated charges

Topic 4:
account savings branch interest rates current high get balance competitive

Topic 5:
banking mobile app atm online support network often limit device
```

Example:

Topic 5: Mobile Banking and Online Services

- **Words:** banking, mobile app, ATM, online, support, network, often, limit, device
- **Insights:** Feedback suggests dissatisfaction with mobile banking applications, ATM services, online services etc.
- **Recommendation:** Improving the functionality and reliability of digital banking services could address these issues.

Generative AI for Retention Strategies

- **Customized Retention Strategies Based On Customer Profiles And Feedback.**
  - **Tools:** Hugging Face 'transformers' library with GPT-2 or GPT-3.
  - **Method:**
    - Input Preparation: Integrate customer data, churn risk, feedback, and sentiment analysis result into detailed prompts.
    - Generation: Utilize generative models to produce tailored communication plans and offers.
- **Integrating Generative AI with Predictive Models and NLP:**

Combine Random Forest for churn prediction with NLP for sentiment analysis to enhance Generative AI for personalized retention strategies.

1	**Retention Strategy**
2	profile: Male Credit Card, Churn Risk: Low, Sentiment: Negative, Feedback: My Credit Card is not generating OTP. Suggested Approach: Provide a guide to resolving OTP issues, offer a security review to enhance the credit card experience.
3	profile: Male Current Account, Churn Risk: Low, Sentiment: Negative, Feedback: The Current Account charges are too high. Suggested Approach: Review account fees, offer a fee reduction or a more suitable account plan, and explain the va
4	profile: Female Loans, Churn Risk: High, Sentiment: Negative, Feedback: The loan prepayment charges are too high. Suggested Approach: Introduce a flexible prepayment policy or discount charges, and provide personalized financial advice

System Integration & Deployment

Architecture:

- **Components:** Integrate Predictive modelling, NLP Analysis (Hugging Face), and Generative AI into a unified system.
- **Flow:** Data Collection → Preprocessing → Predictive Analysis → NLP Analysis → Generative AI for Strategy → Output

Deployment Strategy:

- **Cloud-Based**: Utilize platforms like AWS or Azure for scalable and flexible deployment.

Real-Time Data Processing:

- **Data Streaming:** Implement Apache Kafka or Apache Flink for real-time data ingestion.
- **Model Updates:** Use Docker for containerization and CI/CD pipelines to automate model updates.

Scalability Challenges:

- **Challenges:** Handling large datasets, maintaining quick response times, managing growing user demand.
- **Solutions:** Use distributed databases (e.g., MongoDB), implement load balancing and caching (e.g., Redis), and employ auto-scaling and serverless architectures (e.g., AWS Lambda).

Libraries and Versions Used

Libraries used:

1. **Data Handling and Manipulation:**
   - Pandas: 2.0.0
   - NumPy: 1.25.0
2. **Machine Learning and Predictive Modelling:**
   - scikit-learn: 1.3.0
3. **Deep Learning and NLP:**
   - Transformers (Hugging Face): 4.31.0
   - PyTorch: 2.1.0
   - TensorFlow: 2.14.0
4. **Data Visualization:**
   - Matplotlib: 3.8.0
   - Seaborn: 0.13.2
5. **Deployment and Integration:**
   - Docker: 24.0.3
   - Apache Kafka: 3.4.1
6. **Database and Cloud Storage:**
   - MongoDB: 6.0.4
   - AWS SDK (boto3): 1.26.2

Thank You

Raju Kumar

Computer Science & Engineering

National Institute of Technology, Durgapur

rajukumar6

rajukumar.sde@gmail.com