Project Requirement Document (PRD)

Project Title: Customer Churn Risk Prediction Model

Objective:

The primary objective of this project is to build a **predictive model** that calculates the **Churn Risk Score** for each customer based on their behavioral, demographic, and transactional data. This score will help the business identify customers who are at risk of churning, allowing the company to implement targeted retention strategies. The churn risk score will be represented on a scale from **1 (low churn risk)** to **5 (high churn risk)**.

Dataset Link: https://drive.google.com/drive/folders/1tr6KcaFTTjB1EI58wJhKiC2jR4yLljP-?usp=sharing

Project Description:

Using historical data that encapsulates customer details, purchase patterns, engagement levels, and feedback/complaints, the model will predict the likelihood of churn. This will help the business proactively mitigate churn by engaging with high-risk customers through personalized offers, loyalty programs, or improved service. The project will follow a structured pipeline of data cleaning, exploratory analysis, feature selection, model building, and evaluation.

Data Dictionary:

Feature	Description	Type	Example
`customer_id`	Unique identifier for each customer	Categorical	fffe43004900440036003000 3000
`Name`	Name of the customer	Categorical	Pattie Morrisey
`age`	Age of the customer	Numerical	18
`gender`	Gender of the customer	Categorical	F
`security_no`	Encrypted security number	Categorical	XW0DQ7H
`region_category`	Category of the region (City, Town, Village)	Categorical	Village
`membership_category`	Membership level of the customer	Categorical	Platinum Membership
`joining_date`	Date the customer joined the platform	Date	17-08-2017
`joined_through_referral`	Whether the customer joined through referral (Yes/No/?)	Categorical	No
`referral_id`	Referral ID	Categorical	CID21329

Feature	Description	Type	Example
`preferred_offer_types`	Customer’s preferred offer types (Vouchers, Cards, etc.)	Categorical	Gift Vouchers
`medium_of_operation`	Platform used by the customer (Desktop/Smartphone/Other)	Categorical	Desktop
`internet_option`	Internet option used (Wi-Fi, Mobile Data, Fiber Optic)	Categorical	Wi-Fi
`last_visit_time`	Timestamp of the customer’s last visit	Timestamp	16.08.02
`days_since_last_login`	Number of days since the last login	Numerical	17
`avg_time_spent`	Average time spent on the platform in minutes	Numerical	300.63
`avg_transaction_value`	Average transaction value	Numerical	53005.25
`avg_frequency_login_days`	Average frequency of logins in days	Numerical	17
`points_in_wallet`	Number of reward points in the customer's wallet	Numerical	781.75
`used_special_discount`	Whether the customer has used special discounts (Yes/No)	Categorical	Yes
`offer_application_preference`	Whether the customer prefers offers applied automatically	Categorical	Yes
`past_complaint`	Whether the customer has filed complaints (Yes/No)	Categorical	No
`complaint_status`	Status of customer complaints (Solved, Unsolved, etc.)	Categorical	Solved
`feedback`	Customer’s feedback about the service	Categorical	Products always in Stock
`churn_risk_score`	Target variable representing the churn risk (1-5 scale)	Numerical	2

Key Assumptions:

1. **Churn Risk Score**: The churn risk score ranges from **1 to 5**, with higher scores indicating higher chances of churn.
2. **Customer Interaction Features**: Variables such as `avg_time_spent`, `days_since_last_login`, and `avg_transaction_value` are assumed to strongly correlate with customer engagement and, by extension, the likelihood of churn.
3. **Offers and Complaints**: Features like `preferred_offer_types`, `used_special_discount`, and `past_complaint` reflect customer satisfaction and behavior, which could influence churn risk.

4. **Membership and Regional Influence**: It is assumed that membership tier (`membership_category`) and geographical classification (`region_category`) might influence a customer's retention rate.

Project Steps & Marking Breakdown:

This project is structured into **7 distinct steps**, with a total score of **100 marks**. Each step contributes to building and refining the predictive model and generating business insights from it.

Step 1: Data Cleaning & Preprocessing (20 Marks)

- **1.1 Handling Missing Data** (5 Marks)
  - Identify and impute missing data in relevant fields like `joined_through_referral`, `region_category`, etc.
  - Appropriate handling methods, such as median imputation for numerical data and mode or "Unknown" for categorical data.
- **1.2 Data Type Correction** (3 Marks)
  - Correct data types for fields like `joining_date` and `last_visit_time`.
- **1.3 Encoding Categorical Variables** (5 Marks)
  - Convert categorical variables (`gender`, `region_category`, etc.) into numerical format using One-Hot Encoding or Label Encoding.
- **1.4 Outlier Detection & Handling** (5 Marks)
  - Detect and handle outliers for variables like `avg_transaction_value`, `points_in_wallet` using techniques like IQR or Z-score.
- **1.5 Feature Engineering** (2 Marks)
  - Derive new features, such as `customer_tenure` (from `joining_date`) and engagement-based features from `days_since_last_login`.

Marking Criteria:

- Correct handling of missing data: **5 Marks**
- Data types and encoding handled appropriately: **8 Marks**
- Outliers and feature engineering performed effectively: **7 Marks**

Step 2: Exploratory Data Analysis (EDA) (15 Marks)

- **2.1 Statistical Summaries** (5 Marks)
  - Summarize key numerical variables (`avg_time_spent`, `avg_transaction_value`) with mean, median, and standard deviation.
- **2.2 Visualizations** (5 Marks)
  - Create bar plots, histograms, and box plots for `age`, `churn_risk_score`, `avg_time_spent`, and other relevant features.
  - Plot correlation heatmap to identify relationships between variables.
- **2.3 Customer Segmentation Analysis** (5 Marks)
  - Segment the customers based on `membership_category`, `region_category`, and `gender` to understand churn patterns.

Marking Criteria:

- Comprehensive statistical analysis: **5 Marks**
- Effective use of visualizations for analysis: **5 Marks**

- Meaningful customer segmentation insights: **5 Marks**

Step 3: Feature Selection and Data Splitting (10 Marks)

- **3.1 Feature Selection** (5 Marks)
  - Apply feature selection techniques, such as correlation analysis, feature importance scores, or domain knowledge, to select relevant features.
- **3.2 Data Splitting** (5 Marks)
  - Split the dataset into training and test sets (e.g., 80:20 or 70:30).
  - Stratify the target variable (`churn_risk_score`) if necessary to maintain class distribution.

Marking Criteria:

- Correct and justified feature selection: **5 Marks**
- Appropriate dataset splitting and stratification: **5 Marks**

Step 4: Model Building (25 Marks)

- **4.1 Algorithm Selection** (5 Marks)
  - Select algorithms like Decision Trees, Random Forest, or Gradient Boosting, with clear justification based on the problem and dataset.
- **4.2 Model Training** (10 Marks)
  - Train at least two models on the training dataset and compare their performance.
- **4.3 Hyperparameter Tuning** (5 Marks)
  - Perform hyperparameter tuning (e.g., using GridSearchCV or RandomizedSearchCV) for at least one model to optimize performance.
- **4.4 Cross-validation** (5 Marks)
  - Use k-fold cross-validation (e.g., k=5 or k=10) to ensure model robustness and minimize overfitting.

Marking Criteria:

- Quality of algorithm selection and training: **5 Marks**
- Effective hyperparameter tuning: **10 Marks**
- Cross-validation and model comparison: **10 Marks**

Step 5: Model Evaluation (15 Marks)

- **5.1 Metrics Calculation** (10 Marks)
  - Evaluate models using accuracy, precision, recall, F1-score, confusion matrix, and, if applicable, ROC-AUC score.
- **5.2 Final Model Choice** (5 Marks)
  - Choose the final model based on the evaluation metrics and business needs.

Marking Criteria:

- Appropriate use of evaluation metrics: **10 Marks**
- Justification of the final model selection: **5 Marks**

Step 6: Business Insights & Recommendations (10 Marks)

- **6.1 Insight Generation** (5 Marks)
  - Generate actionable insights based on feature importance and model output.
- **6.2 Retention Strategy** (5 Marks)
  - Suggest targeted retention strategies based on the churn risk model.

Marking Criteria:

- Business-relevant insights: **5 Marks**
- Practical retention strategies linked to the churn model: **5 Marks**

Step 7: Documentation & Reporting (5 Marks)

- **7.1 Final Report** (5 Marks)
  - Prepare a professional report with all findings, methodology, results, and conclusions.
- **7.2 Code Documentation** (5 Marks)
  - Ensure that the code is well-organized, commented, and easily understandable.

Total Marks: 100

Expected Deliverables:

- **Churn Prediction Model**: The final machine learning model, along with evaluation metrics.
- **Business Recommendations**: Actionable retention strategies for high-risk customers.
- **Final Report**: A comprehensive document covering the entire project.
- **Codebase**: Well-documented and modular code.