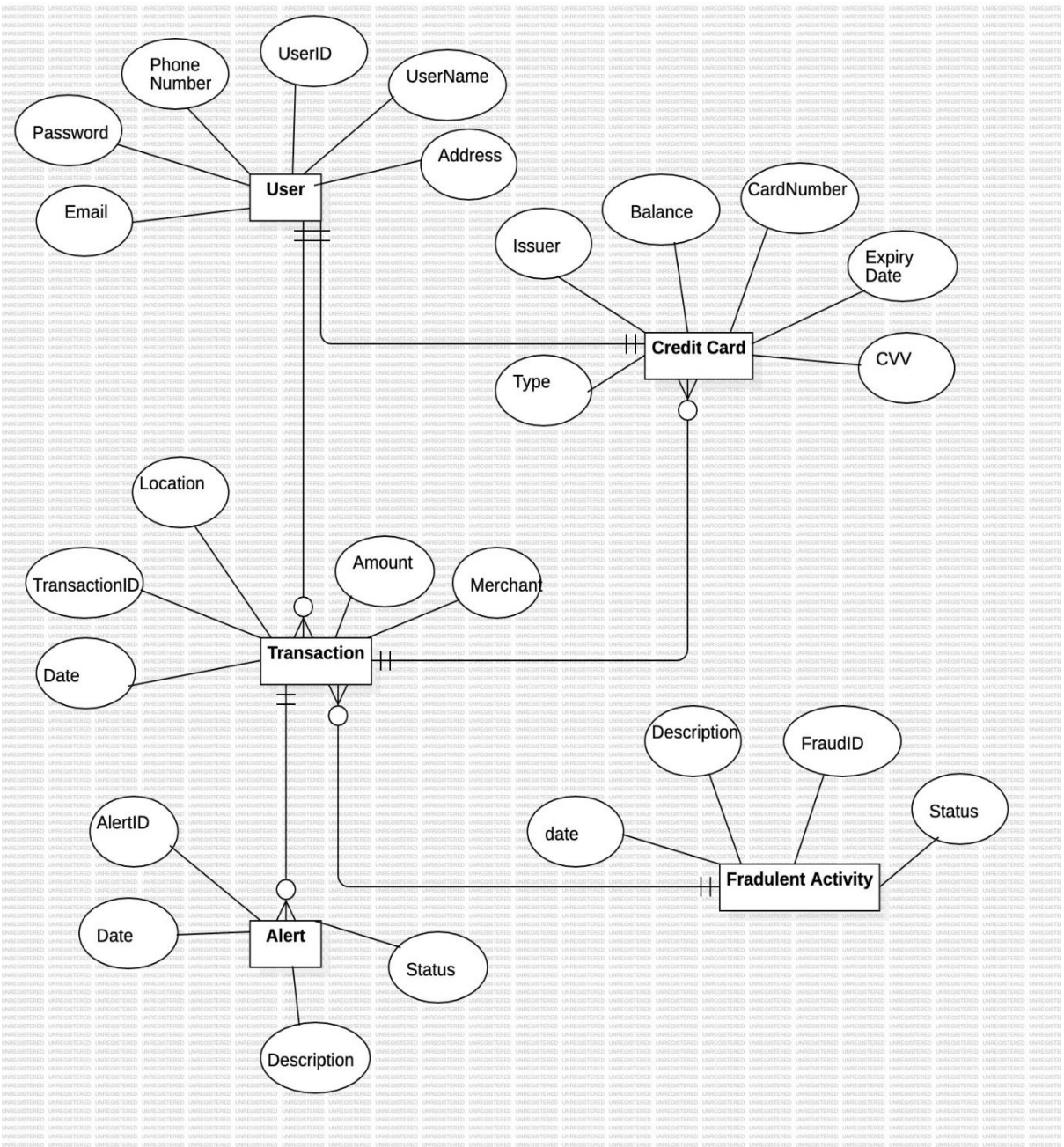
Draw an ER Diagram, And

List the module with a description

Er Diagram



LIST OF MODULES WITH DESCRIPTION

1. Data collection
2. Data pre-processing
3. Data Exploration and data visualization
4. Feature extraction
5. Evaluation model

- Data Collection:

Transaction Logs: These logs contain details about transactions, including the amount, time, and parties involved.

User Account Details: Information about user accounts, such as login history, account age, and user activity.

Location Information: Data related to the geographic location of users during transactions.

Device-Related Data: Information about the devices used for transactions, including device type, operating system, and any unusual behavior.

- Data pre-processing: Pre-processing involves three important steps:

1) Formatting: It is the process of putting the data in a legitimate way so that it is suitable to work with. The format of the data files should be formatted according to the need. The most recommended format is.csv files.

2)Cleaning: Data cleaning is a very important procedure in the path of data science, as it constitutes the major part of the work. It includes removing missing data, reducing complexity with naming categories, and so on.

3)Sampling: This is the technique of analyzing subsets from large datasets, which could provide a better result and help in understanding the behavior and pattern of data.

- Data exploration and data visualization:

1)Class Distributions: Understand the distribution of the target classes (fraudulent vs. non-fraudulent transactions). Imbalanced class distributions can impact model performance, so it's important to be aware of this.

2)Feature Characteristics: Explore the characteristics of individual features. This includes summary statistics (mean, median, standard deviation), identifying outliers, and understanding the range of values.

3)Histograms: Visualize the distribution of numerical features using histograms to identify patterns and anomalies.

4)Scatter Plots: Explore relationships between pairs of features, particularly useful for understanding correlations and potential interactions.

5)Heatmaps: Visualize the correlation matrix of numerical features to identify highly correlated or redundant features.

6)Box Plots: Identify outliers and compare the distribution of features across different classes.

7)Time Series Plots: If dealing with temporal data, visualize trends and patterns over time to identify temporal dependencies.

- Feature extraction:

Dimensionality Reduction: Apply techniques like PCA or t-SNE to reduce the number of features while preserving relevant information.

Feature Selection: Select a subset of the most important features based on statistical tests, feature importance, or domain knowledge.

Transformations: Create new features through mathematical transformations or aggregations of existing features.

Information Retention: Strive to retain as much relevant information as possible during the feature extraction process.

Feature Scaling: Scale features to a standard range to prevent dominance of certain features over others.

Interpretability: Ensure that the extracted features remain interpretable and provide insights into the underlying patterns in the data.

- Evaluation model:

1)Business-Relevant Metrics: Focus on metrics that align with business goals. For fraud detection, precision, recall, and F1 score are crucial, as they directly impact the ability to catch fraudulent activities while minimizing false positives.

2)Cross-Validation for Robustness: Utilize cross-validation to assess the model's robustness by testing its performance across multiple data splits. This helps in obtaining a more reliable estimate of how well the model generalizes to unseen data.

3)Threshold Analysis for Decision Making: Conduct a thorough threshold analysis to understand the trade-off between precision and recall. This analysis aids in setting appropriate decision thresholds based on the specific requirements of the business or application.

4)Explainability for Stakeholder Confidence: Incorporate explainability techniques like SHAP values or LIME to provide transparent insights into how the model arrives at specific predictions. This enhances stakeholder confidence in the model's decisions and fosters trust.

5)Business Impact Visualization for Decision Support: Translate model performance into tangible business impact metrics, demonstrating how the model's predictions contribute to achieving business objectives. This facilitates decision-making by showcasing the real-world value of the model.