三、研究方法

A. 时间序列分析

在数据预处理阶段，我们把数据集中的时间戳转化为便于分析使用的日期格式。然后在数据可视化阶段，我们借助时间序列分析的方法初步探索客户交易行为和时间的关系。通过绘制不同时间维度的客户交易图，我们可以观察到所有账户在总体上的交易趋势，发现可能的周期性变化。

B. 独热编码

在本研究中，为了适当处理分类变量并使其能够被机器学习模型接受，采用了独热编码（One-Hot Encoding）技术。独热编码通过为每个类别生成一个二进制列来转换分类特征，其中1表示特定记录属于该类别，0则表示不属于。这种方法确保了模型能够在不引入可能的数值误解的情况下，正确理解每个类别的独立贡献。在交易金额的预测中，我们没有对全部变量进行独热编码，而是有选择的进行。我们进行了多次尝试，最终选择了表现好的组合。

C. 交易金额预测

在本部分，我们使用了多种机器学习模型来进行交易金额的预测。

**支持向量机(SVR):** 支持向量机用于回归问题称为SVR。SVR的目标是找到一个函数，使得该函数在ε偏差范围内尽可能地接近所有训练样本点。SVR的公式可以表示为：

其中，C 是正则化参数， 是松弛变量，用于处理不完全满足间隔要求的数据点。

**决策树和随机森林:** 决策树通过递归地将数据分割成越来越小的子集来构建预测模型。随机森林则通过结合多个决策树的预测结果来提高预测精度和控制过拟合。它在不同的数据子集上训练多个决策树，每个树的预测结果被平均或以其他方式合成为最终结果。 图提供了一个全面直观的随机森林回归模型的工作原理

**线性回归**：线性回归是预测数值型响应变量的一种方法，通过拟合一个最小化预测误差的线性方程。其模型形式为：

其中，是系数，是特征变量，ϵ 是误差项。

**多层感知机（MLP）**：MLP是一种前馈人工神经网络，它通过一个或多个中间层（隐藏层）对输入特征进行变换和组合，能够捕捉输入变量之间的非线性关系，适用于复杂的回归任务。

每个模型的性能都通过均方误差（MSE）和决定系数（R2）进行评估，以确定哪个模型最适合我们的数据。

D. 用户行为分析

在我们的项目中，为了分析用户的消费行为，我们采用了K-means聚类算法对用户数据进行分组。K-means是一种无监督学习方法，用于将数据点划分到预定数量的簇中，以便将具有相似特征的数据点归为一组。该方法的基本步骤包括选择簇的数量 K 并随机初始化 K 个簇心。然后，算法迭代执行以下步骤直到收敛：

分配步骤：对于每个数据点，根据其特征的欧氏距离，将其分配到最近的簇中心。

更新步骤：更新每个簇的中心，使其成为簇内所有点的均值。

公式上，簇内的每个点 x 被分配到距离最近的簇中心

是该簇的中心。

A. Time Series Analysis

In the data preprocessing phase, we convert the time stamps in the data set into a date format that is easy to use for analysis. Then in the data visualization stage, we use the method of time series analysis to initially explore the relationship between customer trading behavior and time. By plotting customer transactions in different time dimensions, we can observe the overall trading trend of all accounts and discover possible cyclical changes.

B. One-Hot Encoding

In this study, in order to properly handle categorical variables and make them acceptable to machine learning models, a One-Hot Encoding method is used. This method transforms classification characteristics by generating a binary column for each category, where 1 indicates that a particular record belongs to that category and 0 indicates that it does not. This approach ensures that the model correctly understands the independent contributions of each category without introducing possible numerical misinterpretations. In the prediction of transaction amount, we did not carry out unique thermal coding for all variables, but carried out selectively. We tried many times and finally chose the combination that performed well.

C. Transaction Amount Prediction

In this section, we used multiple machine learning models to predict transaction amounts.

**Support Vector Regression (SVR)** is the application of Support Vector Machine (SVM) to regression problems. The goal of SVR is to find a function that approximates the target values as closely as possible while allowing for some deviations for particular data points. The optimization problem of SVR can be expressed as:

图示, 示意图

描述已自动生成

Here, C is the penalty parameter, are slack variables that measure the degree of violation for the data points that do not fall within the established margin.

**Decision Tree Model** constructs predictive models by recursively partitioning the data into increasingly smaller subsets. \textbf{Random Forest Model} improves prediction accuracy and control overfitting by combining the predictions of multiple decision trees. They train numerous decision trees on different subsets of data, where the prediction of each tree is averaged or otherwise synthesized into the final outcome. Fig.\ref{fig:randomforest} provides an intuitive illustration of how a Random Forest Regression Model works.

**Linear Regression** is a method for predicting a numerical response variable by fitting a linear equation that minimizes the prediction error. The model form is

**Multilayer Perceptron (MLP)** is a type of feedforward artificial neural network that transforms and combines input features through one or more intermediary layers (hidden layers). It is capable of capturing nonlinear relationships between input variables and is suitable for complex regression tasks.

The performance of each model is evaluated using the Mean Squared Error (MSE) and the Coefficient of Determination (R2) to determine which model is the best fit for our data.

D. User Behavior Analysis

In our project, to analyze consumer behavior, we utilized the K-means clustering algorithm to group user data. K-means is an unsupervised learning method that classifies data through the following steps:

* Initialization: Choose the number of clusters \(K\) and randomly establish \(K\) cluster centers
* Assignment: Assign each data point to the nearest cluster center based on the Euclidean distance between the point and the cluster center.
* Update: Update the center of each cluster to be the mean of all points assigned to the cluster.

This process is repeated until the cluster centers stabilize or a predetermined number of iterations is reached. By analyzing features such as transaction amounts, frequency, and account balances, we employ K-means to identify different user groups. This approach helps us understand the consumption patterns and preferences of various groups.