**Exploratory analysis of data**

According to the knot analysis, it can be found that the single transaction amount is basically concentrated in the range of 0-1000, and the frequency of the last transaction amount of different users from high to low becomes a steady decreasing trend, but the frequency of the time from the last transaction time has a tendency to rebound with the increase of time, which indicates that the users are mainly concentrated in the two ends of active and inactive intervals. Subsequently, by looking at the covariance matrix of the independent variables, it can be seen that the correlation between the two variables is almost zero, except for the current transaction amount, which seems to be somewhat correlated with the last transaction amount. the distribution of the number of transactions occurring tends to be more even in terms of device and type of product, and geographically Houston is the most prominent, while New York has the least number of transactions.

**Hypothesis**

The box plot shows that the fraudulent types of transactions are significantly higher than normal transactions, so it is reasonable to assume that there is some correlation between the type of transaction and the amount of transactions, and since during data exploration it was found that the difference in the time of the current transaction from the time of the last transaction varies more drastically, the time factor is also added into consideration, i.e., the type of transaction is related to the cumulative amount of transactions.

**Data preparation**

Firstly the dataset is checked for missing values and through the results it can be found that there are no missing values. Subsequently the transaction occurrence time is converted from object type to pandas time type

**Feature engineering**

Three new attributes are introduced based on the existing data, the cumulative number of transactions, the cumulative transaction amount, and the cumulative average transaction amount for this user as of the current time in the dataset time period.

It is clear that the transaction ID is not associated with being fraudulent, and it is removed to reduce data complexity.

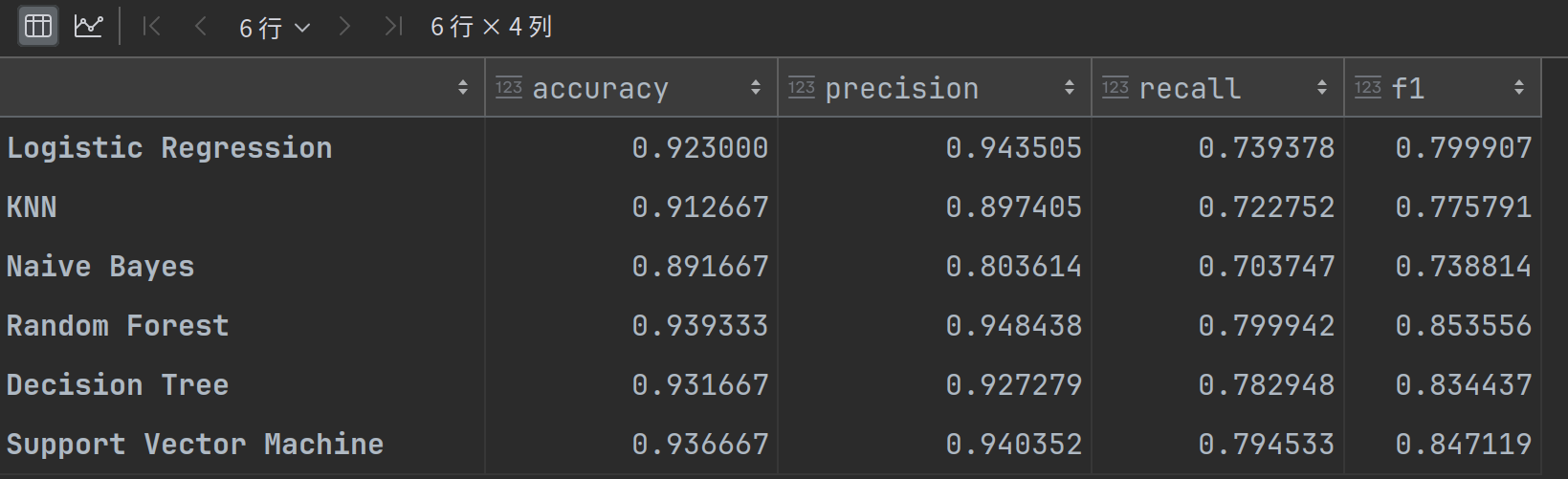
The type of commodity traded, the region in which the transaction is made, and the device on which the transaction occurs are uniquely and thermally encoded so that the machine learning model can use the information implicit in them.

**Model building and optimization**

In this experiment, a total of six models were defined and range model parameters were set to optimize the models with hyperparameters. These models include logistic regression (penalty parameters of l1 and l2, C parameters ranging from 0.01, 0.1, 1, 10, and 100, and solver as liblinear), K-nearest neighbor (n\_neighbors parameters of 5, 7, 9, 15, 35, 45, and 55, weights of uniform and distance, the metric as euclidean, manhattan, and minkowski), plain Bayes (var\_smoothing parameters as 1e-9, 1e-8, and 1e-7), random forest (n\_estimators parameters as 50, 100, and 200, max\_features as sqrt and log2, and max\_depth as None, 10, 20 and 30), decision trees (criterion parameters as gini and entropy, max\_depth as None, 10, 20 and 30, min\_samples\_split as 2, 5 and 10, and min\_samples\_leaf as 1, 2 and 4) and support Vector Machine (C parameters of 0.1, 1, 10 and 100).

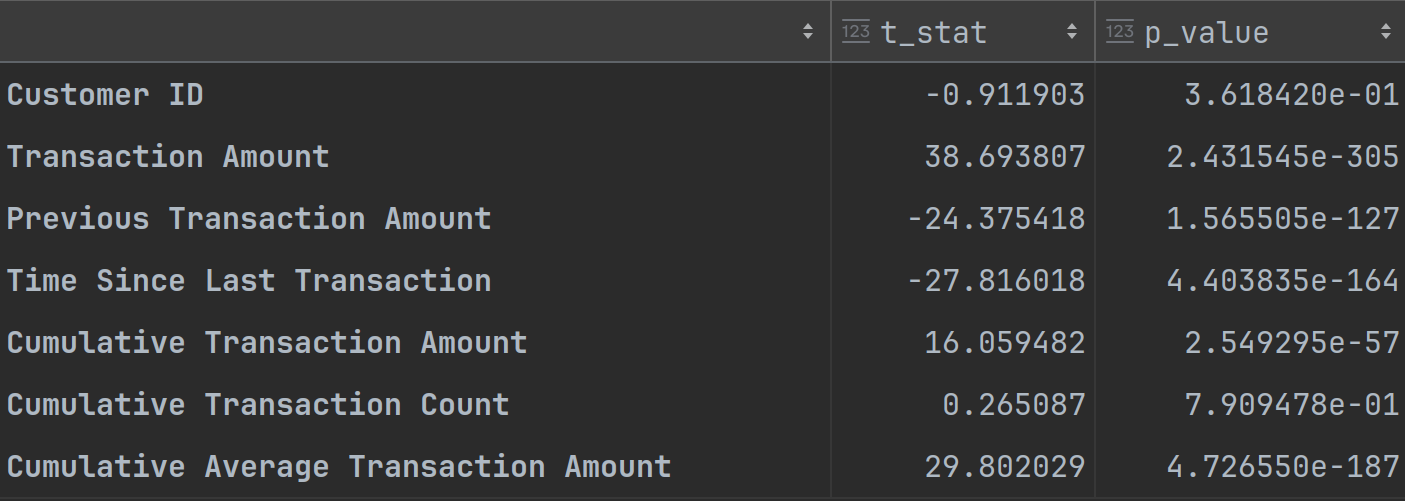
**Model Evaluation**

The results obtained through grid search as well as cross validation are as follows.

The results show that the random forest model performs the best on all metrics with the highest accuracy (0.939333), precision (0.948438), recall (0.799942), and F1 score (0.853556), followed by the support vector machine and decision tree models, and these results validate our hypotheses that the characteristics of transaction amount, frequency of transactions, and time intervals are crucial for identifying fraudulent transactions. The excellent performance of the random forest model shows that it can effectively balance false positives and false negatives and has high practical value in real business applications.

**Analysis of results**

Comparison of each numerical feature of fraudulent and non-fraudulent transactions by t-test gives the following results：



Transaction amount, last transaction amount, time since last transaction, cumulative transaction amount and cumulative average transaction amount can be found to be significantly different between the two groups, indicating that these features play an important role in identifying fraudulent transactions. Whereas, customer ID and cumulative number of transactions are not significantly different between the two groups, thus they have less impact on fraud detection. These findings validate the initial hypothesis that transaction amount, transaction frequency and time interval are the key features in identifying fraudulent transactions.

In addition, since Random Senri's model performs best, it can be used to identify transaction types.

Enterprises can integrate our fraud detection model into their transaction processing systems, generating fraud probability scores by analyzing the characteristics of each transaction in real time (transaction amount, transaction frequency and time interval, etc.), automatically determining and blocking potential fraudulent behavior, and at the same time, reviewing and analyzing the historical transaction data through regular batch processing, to ensure the security of the transaction and customer trust.