CourseWork2: Find suspect trades

Section 1: Introduction

In the finance market, traders might make mistakes when submitting their trades, and the mistake trades data may cause problems. With a case of finding out mistake trades data, the report will discuss some methods of recognizing mistake trades and implement some of them.

The report is structured as follows: the first section briefly introduces the case, and the second section presents the solution to find out mistake trades. In the fourth and fifth sections, the report discusses several approaches deeply. Implementation and results are in the sixth and seventh sections. Lastly, the report gives a conclusion.

Section 2: **Present business case, challenges and data**

Background

Traders submit single trades at any point in time during the day of 2021-11-11 & 2021-11-12. Traders might make mistakes when they submit their trades, it is your responsibility to verify that trades submitted by Front Office are genuine and in line with market expectations. If a trade is deemed to be suspect, (i.e. incorrect trade price or incorrect quantity), it must be reported.

challenges:

Find out suspect trades between 2021-11-11 and 2021 –11 -12.

Data introduction:

Dataset	Introduction
Data_trades [mongodb]	Data of trades
Equity_prices	Data of equity prices
Trades_suspects	Data of suspect trades. This is our output
Portfolio_positions	Data of portfolio positions

Section 3: **Present the solution to the challenge:**

In this case, we define "suspects trades" as trades with abnormality. And the abnormality may be caused by the following reasons. For each reason, we have corresponding solutions:

Abnormality	Cause	Solution
Inconsistency & Not genuine	Fat finger error.	Matching datasets and cross-validation
	Mis-pricing.	Outlier detection on Time series.
	Manipulating.	Outlier detection
		Visualization

More explanations about causes and solutions are following:

Fat finger error & Matching datasets and cross-validation

Fat finger error: A fat finger error is a human error caused by pressing the wrong key when using a computer to input data.

Solution: In this case, several datasets come from different sources. Thus, it's possible to do dataset cross-validation—for example, matching price calculated from trades dataset with price from equity dataset, to check the consistency.

Manipulating & Outlier detection

Manipulating: Trades with abnormal data may have the suspension to be manipulated. These trades are not genuine. For example, a trader suddenly had an unusually large transaction. Another example is continuous trading at a specific time.

Solution: Outlier detection methods could be the solution to the challenge. Because some suspect trades have abnormal value among datasets. For example, fat-finger errors may make the price (or nominal) outstanding on the scale, which is extremely high or extremely low among datasets. Thus, detecting outliers can help to find out suspect trades.

Mispricing & Outlier detection on Time series.

Mispricing: Mispricing leads to trades containing the obvious abnormal price, for example, a trading price not consistent with the trend, moment, even the volatility of history stock price series.

Solution: By applying outlier detection on time series like stock price series. We can capture these abnormalities.

Visualization

Visualization like scatter plots can help us observe the data distribution. It may help us to find outliers directly. Besides, data exploring via visualization not only gives us an intuitive understanding of the dataset's features, but also provides direction for further detecting and choosing appropriate models.

Section 4: Model selection about outlier detection

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Here exists several different kinds of outlier detection methods in the literature.

https://blog.csdn.net/weixin_26739079/article/details/109123172

One type of outlier detection method is the statistic approach. According to Fauconnier, C., and G. Haesbroeck. (2009), Minimum Covariance Determinant (MCD) could be used to detect outliers. MCD searches for the subset of a specified number of data points whose covariance matrix contains the lowest determinant. PCA allows orthogonal transformation, which helps to distinguish outliers and normal observations. (Hussain, Saddam, Mohd Wazir Mustafa, Touqeer Ahmed Jumani, Shadi Khan Baloch, and Muhammad Salman Saeed. 2020)

Clustering combined with distance-based metrics and density-based metrics are also common in detecting outliers. Aims to detect anomalous data points with the help of the density of its encircled space that is detached by regions of low-density observations.

Distance-based approach and density-based approach is also very popular for detecting outliers. The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) aims to detect anomalous data points with the help of the density of its encircled space that is detached by regions of low-density observations (Li, Shi-wu, Yi Xu, Wen-cai Sun, Zhi-fa Yang, Lin-hong Wang, Meng Chai, and Xue-xin Wei. 2015).

Besides, machine learning method can be applied in detecting outliers. Isolation Forest is based on the idea that outliers represent data points that are few and different. And like any tree ensemble methodology, it is based on decision trees. (Santosuosso, U., A. Cini, and A. Papini. 2020).

Sumarise of Outlier detection models:

Type	Model	Limitations
Statistic Approach	MCD	Assume data has Gaussian distribution.
	PCA	Lack of ability to dealing with non-Gaussian distribution data.
Distance-based Approach	K-means Local Outlier Factor	Not robust to scale. Requires A threshold is to determine outliers, but it's not easy to find a meaningful threshold.
Density-based Approach	DBSCAN	
Isolation-based Approach	Isolation Forest	The desired effect may not be achieved on a small sample.
Classification-based Approach	One-Class SVM	Not suitable for large sample models due to computation consumption.

Section 5: Which Approaches is used and why

This report carries out the process of outlier detection as following:

Step1: With interactive visualization, have a preliminary understanding of the data.

Step2: Using matching datasets and cross-validation to find consistency between trades dataset and equity dataset.

Step3: Applying Isolation Forest to find outliers, in other words, suspect trades. There are two reasons for choosing Isolation Forest. The first is the large sample without Gaussian distribution. The second reason is that the data has high dimensions (many features). Thus, according to the analysis of limitations of several outlier detection models, Isolation Forest should be the best choice.

Section 6: Implementation

In the project use Python in implementation. And the whole project contains following scripts:

Name	File Type	Introduction
ReadMe	Markdown	Introduction
Main.py	Python script	Run this script to start the project.

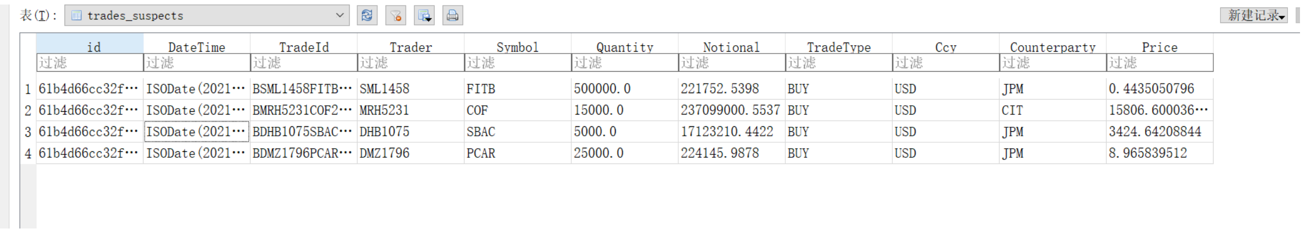
Config	Folder	
Script_params.py	Python script	Set params
Script.config	Config document	Set config
moduls	Folder	
CV	Folder	
match.py	Python script	Implement matching datasets & cross-validation
db	Folder	
CreateTable.py	Python script	After getting suspect trades data, this script can create a table in sql database ("Equity.db").
Equity.db	sql database	Output data
Equity_original.db	Equity.db	A copy of input data
Undate_position.py	Python script	Update "equity_position" table in "Equity.db"
IF	Folder	
Isolation_forest	Python script	Implement outlier detection via Isolation Forest
Visualization	Folder	
dash_plot.py	Python script	Implement interactive visualization via dash
Static	Folder	
Test	Folder	

Tips 1: Except for Main.py, every script is encapsulated in a function.

Tips 2: Dash is an open source library released under the permissive MIT license. It's design for implementing interactive visualization. More details can be found in: https://dash.plotly.com/introduction

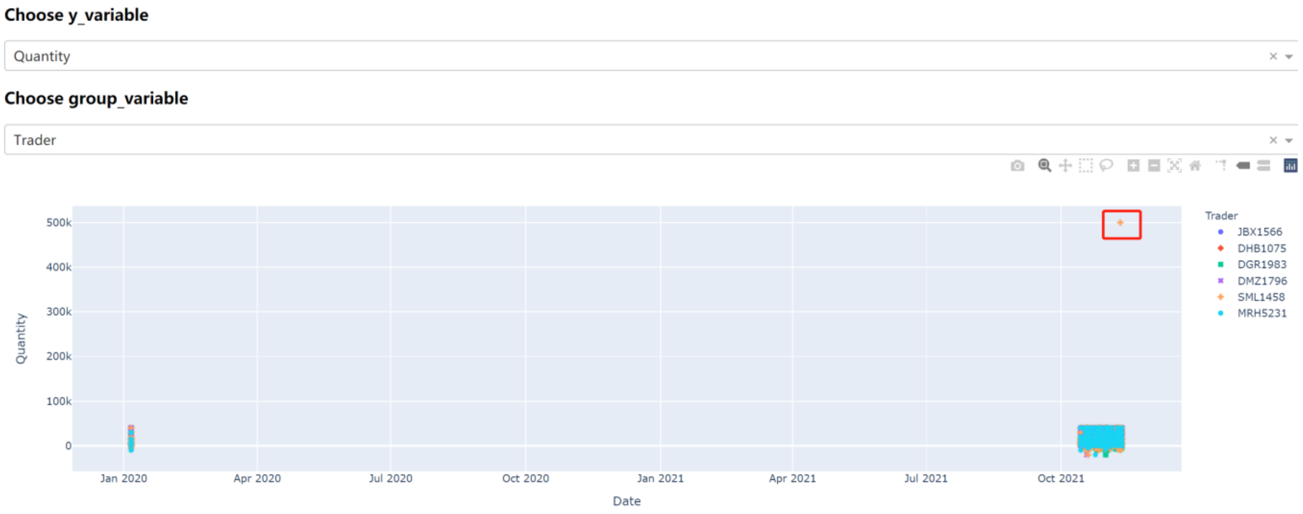
Section 7: Results

4 suspect trades were founded. They are obvious outliers. And they are not consist with data from equity_price.



	id	DateTime	TradeId	Trader	Symbol	Quantity	Notional	TradeType	Ccy	Counterparty	Price
	过滤	过滤	过滤	过滤	过滤	过滤	过滤	过滤	过滤	过滤	过滤
1	61b4d66cc32f…	ISODate(2021…	BSML1458FITB…	SML1458	FITB	500000.0	221752.5398	BUY	USD	JPM	0.4435050796
2	61b4d66cc32f…	ISODate(2021…	BMRH5231COF2…	MRH5231	COF	15000.0	237099000.5537	BUY	USD	CIT	15806.600036…
3	61b4d66cc32f…	ISODate(2021…	BDHB1075SBAC…	DHB1075	SBAC	5000.0	17123210.4422	BUY	USD	JPM	3424.64208844
4	61b4d66cc32f…	ISODate(2021…	BDMZ1796PCAR…	DMZ1796	PCAR	25000.0	224145.9878	BUY	USD	JPM	8.965839512

By visualization, we can check there are abnormal trades.



Choose y_variable

Quantity

Choose group_variable

Trader

Section 8: Conclusion: Conclude the work:

Though it's not easy to find out suspect data, but check consistency and outlier detection are good ways to detect abnormality in trades data.

Reference:

[1] Fauconnier, C., and G. Haesbroeck. "Outliers Detection with the Minimum Covariance Determinant Estimator in Practice." Statistical Methodology 6.4 (2009): 363-79. Web.

[2] Hussain, Saddam, Mohd Wazir Mustafa, Touqeer Ahmed Jumani, Shadi Khan Baloch, and Muhammad Salman Saeed. "A Novel Unsupervised Feature-based Approach for Electricity Theft Detection Using Robust PCA and Outlier Removal Clustering Algorithm." International Transactions on Electrical Energy Systems 30.11 (2020): N/a. Web.

[3] Li, Shi-wu, Yi Xu, Wen-cai Sun, Zhi-fa Yang, Lin-hong Wang, Meng Chai, and Xue-xin Wei. "Driver Fixation Region Division–oriented Clustering Method Based on the Density-based Spatial Clustering of Applications with Noise and the Mathematical Morphology Clustering." Advances in Mechanical Engineering 7.10 (2015): 168781401561242. Web.

[4] Santosuosso, U., A. Cini, and A. Papini. "Tracing Outliers in the Dataset of Drosophila Suzukii Records with the Isolation Forest Method." Journal of Big Data , 7 , Article 14. (2020) (2020): Journal of Big Data , 7 , Article 14. (2020). Web.