cloudwalk

DETECTING ANOMALIES IN FINANCIAL TRANSACTIONS

Business Monitoring Intelligence Analyst Case - Cloudwalk

Rodrigo Mantovani

EXECUTIVE SUMMARY

1. First task

- Exploratory data analysis;
- Custom anomaly score (statistical & machine learning methods);
- Determining anomaly levels.

2. Second task

- Filtering data without anomalies;
- Developing predictive machine learning models for normal behavior;
- Setting thresholds for abnormal behavior;
- Development of anomaly detection system;
- Development of anomaly alert system;
- Integrating everything in a dashboard;

1. First task

The data: checkout of POS data

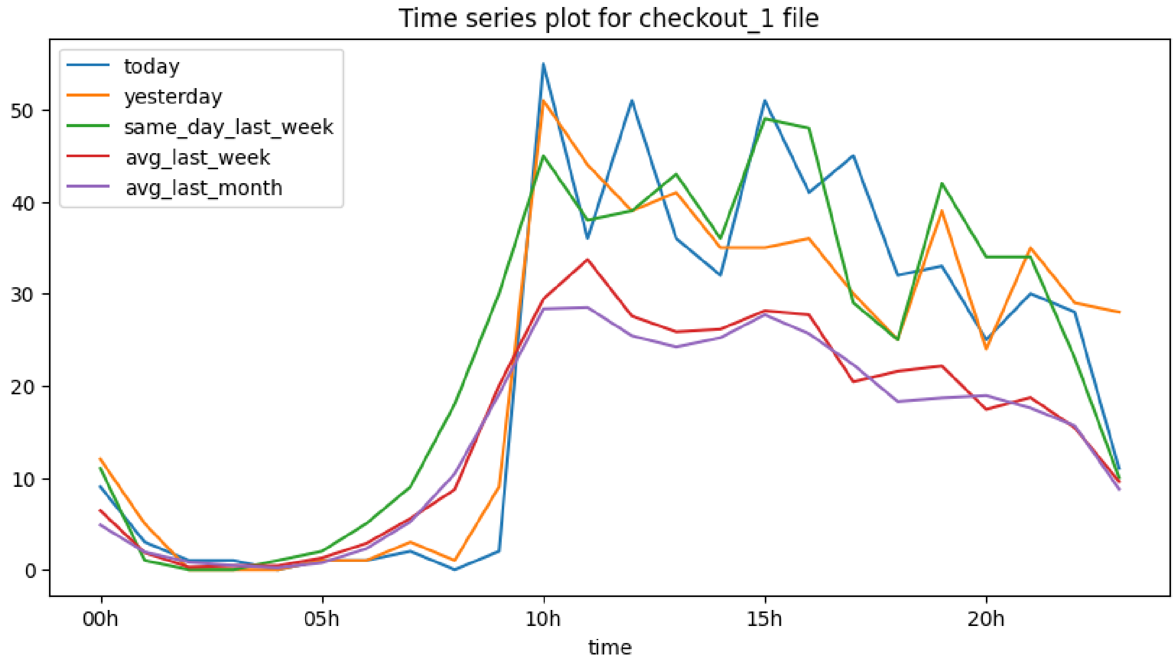
- Contains number of sales by hour, comparing the same values for today, yesterday, the same day last week, the average of the last week, and the average of the last month;
- Two datasets, each ranging from 00h to 23h (one day).

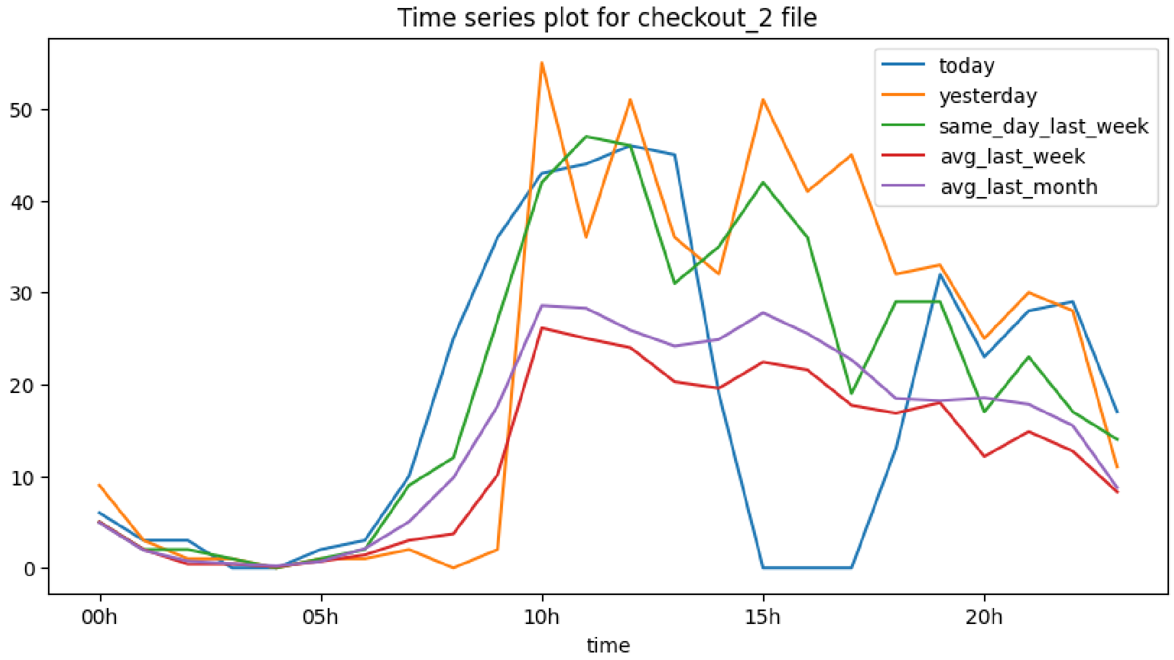
The objective: to identify abnormal behavior referring to "today" in both datasets.

The strategy:

- Conducting Exploratory Data Analysis to better understand the data, its underlying dynamics, and the distribution;
- Computing the deviation between "today" and the other days and averages presented in the dataset;
- Using these deviation values to perform both statistical and machine learning tests for anomaly detection and defining a custom anomaly score based on these tests;
- Determine anomalies based on the anomaly scores;

EXPLORATORY DATA ANALYSIS - VISUALIZATION

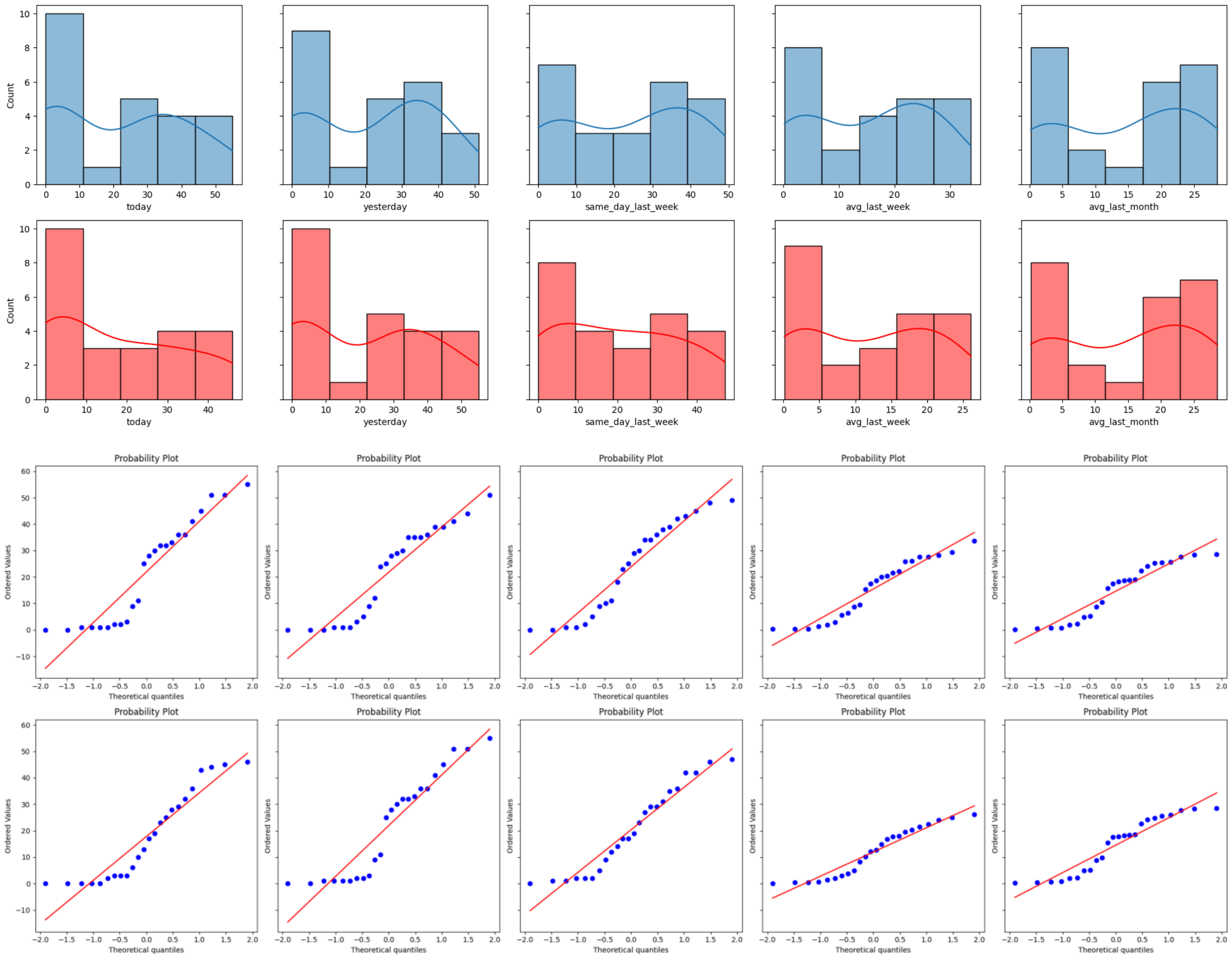




1. First task

EXPLORATORY DATA ANALYSIS - DISTRIBUTION

cloudwalk



Shapiro-Wilk test

Dataset	Statistic	P-value
checkout_1	0.9213	0.0000
checkout_2	0.9129	0.0000

Null-hypothesis rejected. Data appears not to be normally distributed.

1. First task

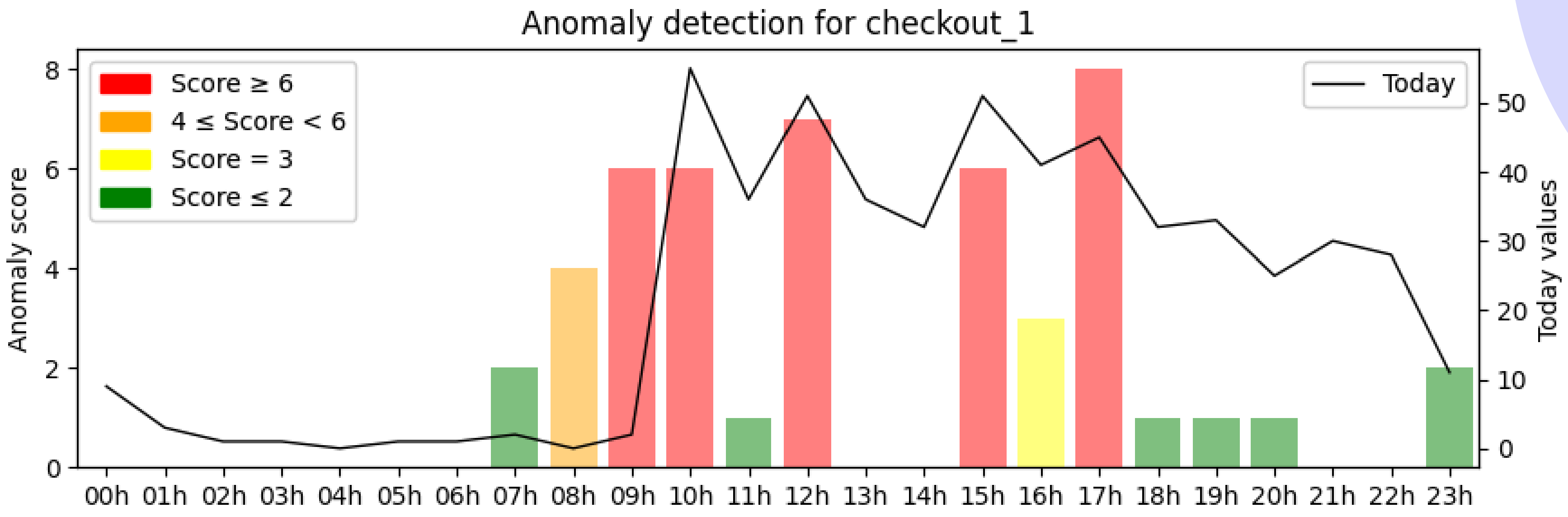
OUTLIER DETECTION

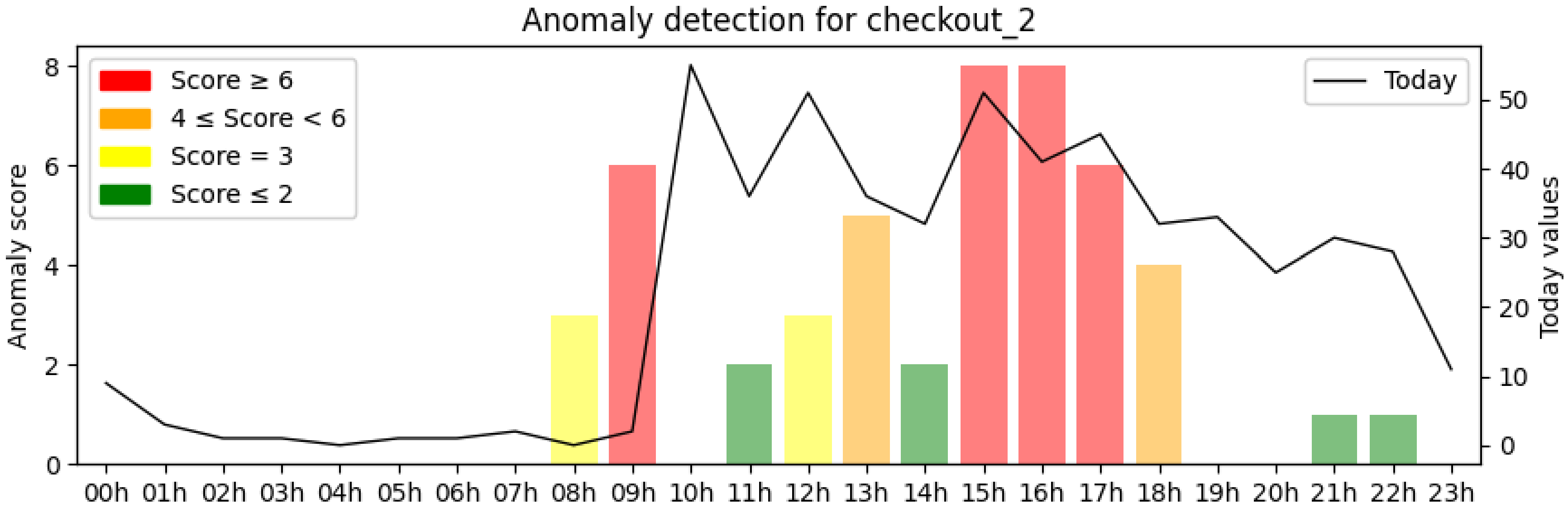
cloudwalk

1. **Compute the deviation** between "today" and the other days and averages presented in the dataset:
   - **Feature engineering:** (today – yesterday), (today – same day of last week), (today – average of last week), (today – average of last month);
2. **Statistical test:** check which deviations exceed 150% of the standard deviation of the column the deviation is based on.
3. **Machine learning test**: use Isolation Forest to determine which data points are most easily isolated from the rest
4. **Custom anomaly score:** since both statistical and ML tests will be performed for each row (hour) of data, 8 different outlier tests will be performed. The anomaly score represents the amount of tests in which each point was flagged as an outlier.

Anomaly score	Anomaly level
Score ≥ 6	Severe
4 ≤ Score < 6	High
Score = 3	Possibly an anomaly
Score ≤ 2	Not an anomaly

OUTLIER DETECTION

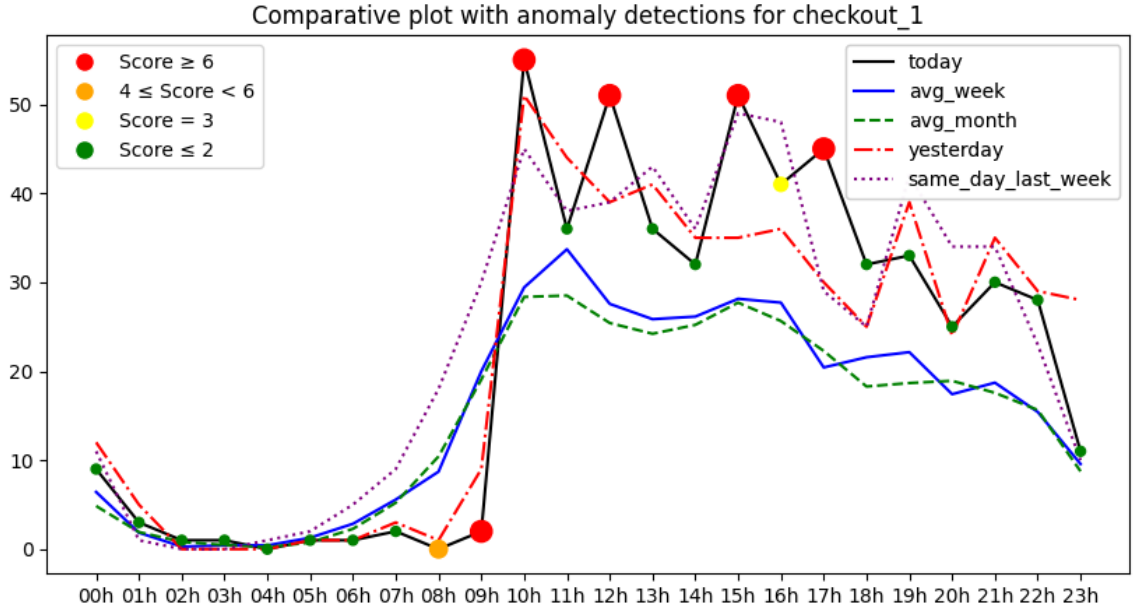


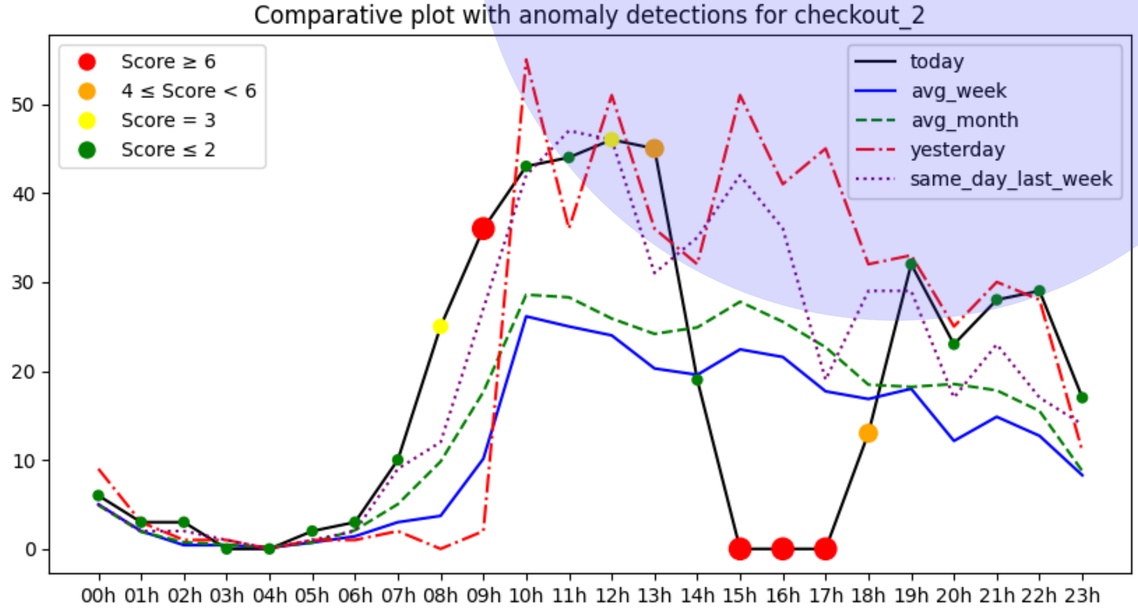


1. First task

TASK CONCLUSION

cloudwalk





time	Anomaly score	Anomaly level
08h	4.0	High
09h	6.0	Severe
10h	6.0	Severe
12h	7.0	Severe
15h	6.0	Severe
17h	8.0	Severe

time	Anomaly score	Anomaly level
09h	6.0	Severe
13h	5.0	High
15h	8.0	Severe
16h	8.0	Severe
17h	6.0	Severe
18h	4.0	High

1. First task

2. Second Task

cloudwalk

The data:

- Contains transaction data by minute, grouping the total transactions in that minute by status: ‘approved’, ‘denied’, ‘reversed’, ‘refunded’, ‘backend_reversed’, ‘processing’, and ‘failed’;
- Two datasets, each ranging from 00h00min to 23h59min (one day);

The objective: to build a monitoring system that receives real time data and alerts in real time in case of abnormal behavior in ‘failed’, ‘reversed’ and ‘denied’ transactions.

The strategy:

- Select the first dataset as a baseline to identify what characterizes abnormal and normal behavior;
- Build a query that organizes the data;
- Use Isolation Forest to determine what points are considered outliers for each of the 3 status we are working with: ‘failed’, ‘reversed’ and ‘denied’.
- Train 3 predictive Random Forest Regressor models, one for each status;
- Select the second dataset as a test dataset, i.e., the subject of anomaly detection;
- Use the deviations between predicted and actual values to determine anomalies;
- Build a system to send automatic alerts in case of anomalies;
- Integrate everything in a dashboard.

PREMISES

cloudwalk

1. The **main objective** is to simulate a real-life monitoring task for anomaly detection;

2. **Data variety**: the system will be built with data from a single day, which would make the models prone to overfitting to particular trends that happened in that day only and don't represent the actual data distribution. However, the objective is to build a consistent structure that would be able to incorporate more data in case it is available, which is very likely in a business setting.

This was taken into account when analyzing results

ORGANIZING THE DATA AND DEFINING BASELINES

cloudwalk

The data includes counts for each status in each minute;

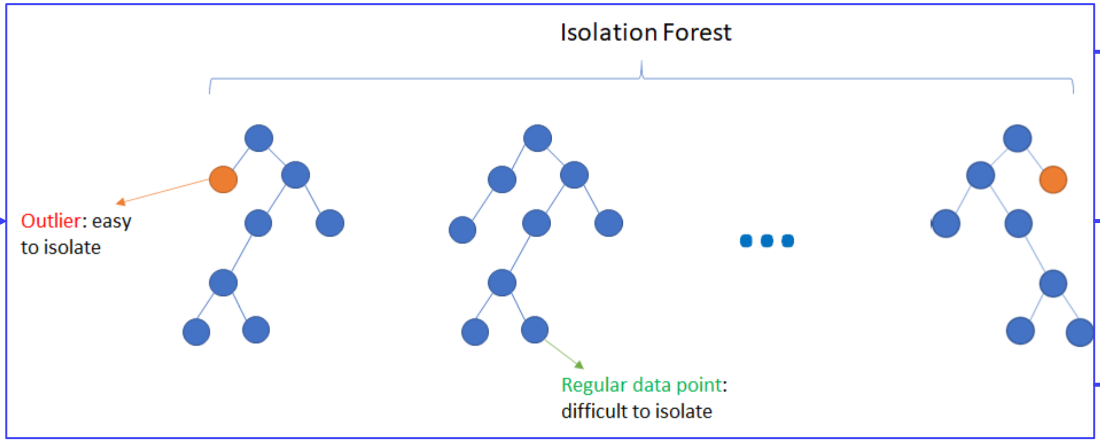
1. Grouping 'backend_reversed' and 'reversed' into a single feature;
2. Transforming the counts of each status into ratios (%), in order to include the effects of seasonality among a single day;
3. Data manipulation: defining continuous time-based features, such as "time of day (in minutes)" and categorical time-based features, such as "time bin" (representing the part of the day in which the transactions occured)

Data with anomalies

"Failed" dataset

"Reversed" dataset

"Denied" dataset



Data without anomalies

"Failed" dataset

"Reversed" dataset

"Denied" dataset

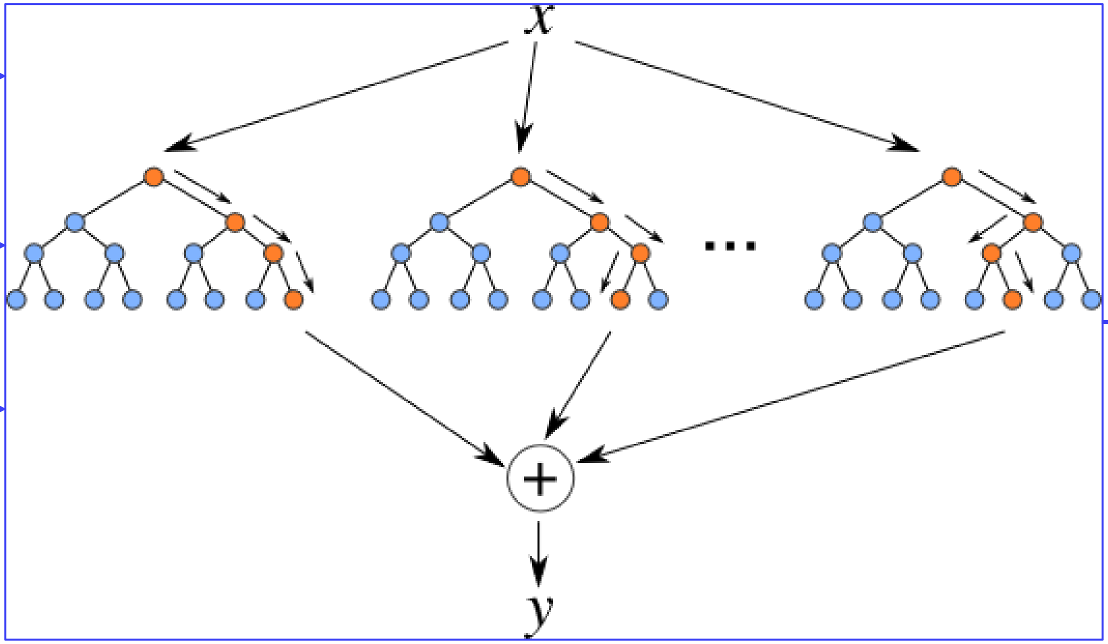
2. Second task

TRAINING PREDICTIVE MODELS AND DEFINING THRESHOLDS

Data without anomalies

- "Failed" dataset
- "Reversed" dataset
- "Denied" dataset

Random Forest Regressors



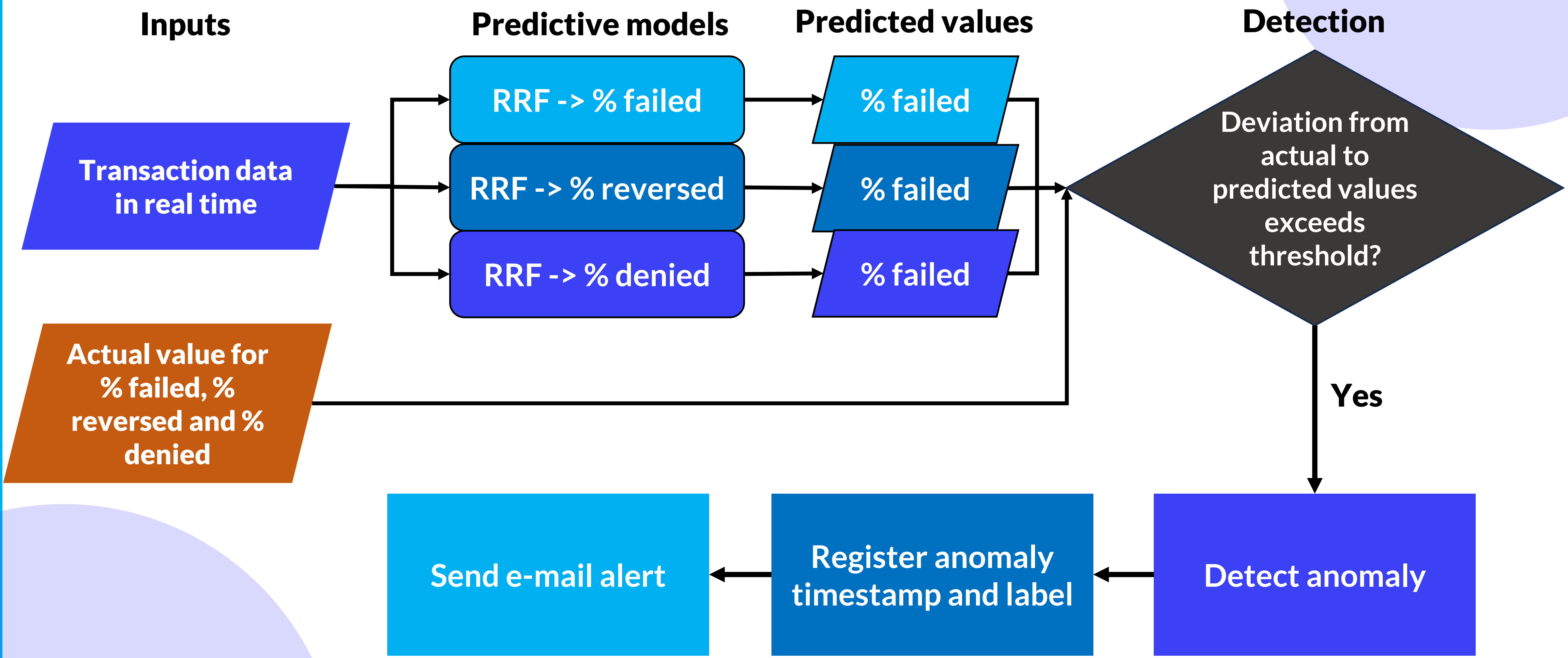
Predictive models:

Predict, in real time, the % of transactions in the desired status

Actual value - Predicted value > **Threshold: 3 x (standard deviation of original data)** → **ANOMALY FLAG**

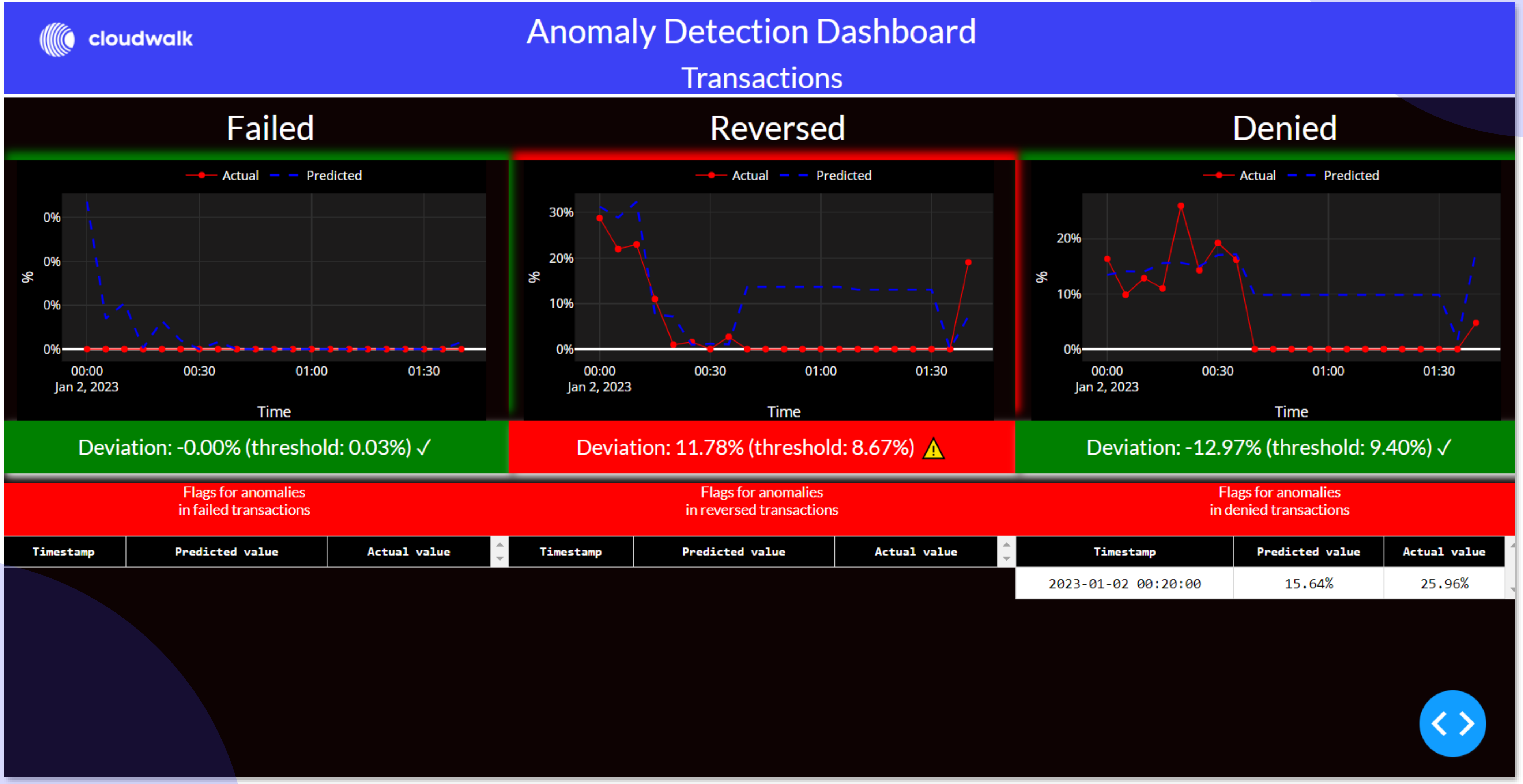
THE SYSTEM

2. Second task



cloudwalk

THE DASHBOARD



cloudwalk

DETECTING ANOMALIES IN FINANCIAL TRANSACTIONS

Business Monitoring Intelligence Analyst Case - Cloudwalk

THANK YOU!

Rodrigo Mantovani