Taxi Order Anomaly Detection from Trip-Level Data

Machine Learning Exploratory Project
Robert Wang UID 01586140
December 2, 2025

1 Introduction

This project aims to detect abnormal taxi orders, particularly detour behaviors, based solely on aggregated trip-level order data rather than trajectory-level GPS data. Abnormal taxi orders typically exhibit inconsistencies between travel distance, duration, and fare, which may indicate fraudulent driving or inefficient routing. Traditional rule-based methods rely on fixed thresholds, such as distance-to-time ratios, and fail to adapt to complex, nonlinear traffic dynamics. Therefore, a data-driven machine learning framework is adopted to automatically learn the statistical regularities of normal trips and identify outliers that deviate significantly from these learned norms.

2 Data Description

2.1 Data Source

The dataset contains ride-hailing trip records collected from the city of Nanjing, China, covering a continuous six-month period starting from April 1, 2023. In total, it includes approximately $180 \times 500{,}000 \approx 90$ million trip records. Each daily dataset consists of around half a million trips, and each record provides origin and destination coordinates, departure and arrival timestamps, total driving distance, and trip fare. To enable regional aggregation and spatial analysis, both the origin and destination points are mapped to administrative districts using a regular city grid system with a spatial resolution of $0.01^\circ \times 0.01^\circ$.

2.2 Raw Attributes

The raw dataset contains eight essential fields directly recorded from the ride-hailing platform. These fields serve as the foundation for feature derivation and subsequent modeling.

Field Name	**Description**	**Unit**	**Example**
DEP_LON	Longitude of trip origin (departure location)	degrees	118702884
DEP_LAT	Latitude of trip origin (departure location)	degrees	32117048
DEP_TIME	Timestamp of trip departure in integer format (`YYYYMMDDhhmmss`)	–	20230401094512
DEST_LON	Longitude of trip destination	degrees	118727194
DEST_LAT	Latitude of trip destination	degrees	32112422
DEST_TIME	Timestamp of trip arrival in integer format (`YYYYMMDDhhmmss`)	–	20230401100034
DRIVE_MILE	Total driving distance of the trip as reported by the meter	km	3.01
PRICE	Total fare charged for the trip	CNY	9.0

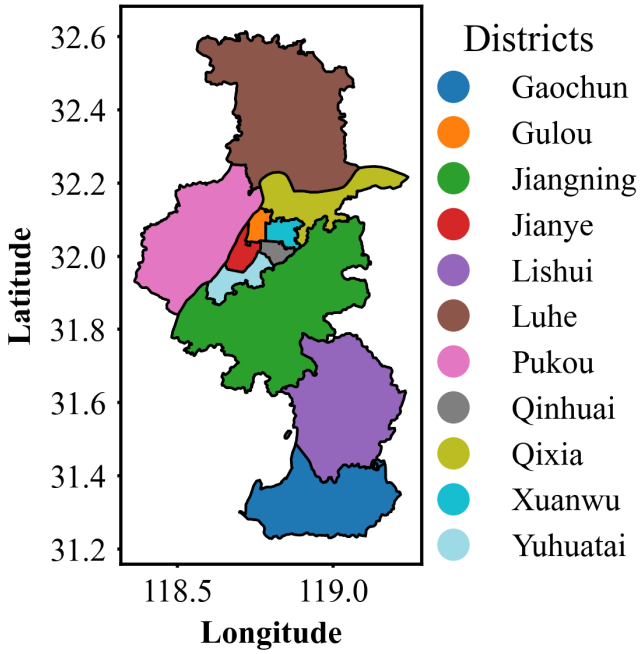

Figure 1: Study area: administrative map of Nanjing City.

Each record represents one completed trip, including departure and arrival times, thus allowing calculation of travel duration, speed, and spatial displacement.

3 Data Cleaning and Preprocessing

The data preparation process follows several key steps:

1. **Duplicate removal:** Trips with identical origin, destination, and time information were removed to avoid repeated entries.
2. **Type conversion:** All numeric fields, such as distance and fare, were converted to floating-point types. Timestamps were transformed from integer format (`YYYYMMDDhhmmss`) into `datetime` objects.
3. **Spatial filtering:** Trips outside the bounding box of Nanjing (longitude 118.36–119.24, latitude 31.23–32.62) were excluded.
4. **Temporal filtering:** Only trips whose departure and arrival both occurred on April 1, 2023 were retained.
5. **Basic validity checks:** Records with nonpositive travel time, nonpositive distance, or nonpositive fare were removed.

After these operations, the cleaned dataset provides a reliable foundation for exploratory analysis and model training.

3.1 Derived Features

Based on the raw fields, the following additional features are constructed:

- **Trip duration** (`DRIVE_TIME`): difference between destination and departure timestamps, expressed in minutes.
- **Trip distance** (`DRIVE_MILE`): total driving distance in kilometers as recorded in the order data (also used as a target variable in some models).
- **Peak-hour indicator** (`IsPeakTime`): binary variable equal to 1 if the order occurred during peak commuting hours (e.g., 7–9 a.m. or 5–7 p.m.), and 0 otherwise.
- **Weekend indicator** (`isWeekend`): equal to 1 if the trip occurred on Saturday or Sunday, and 0 for weekdays.
- **Morning-start indicator** (`isTime`): equal to 1 if the trip began before 8:00 a.m., and 0 otherwise.
- **Trip-length category** (`isLong`): a categorical variable reflecting distance level: short ($\leq$3 km), medium (3–12 km), long (12–24 km), or very long (>24 km).
- **Operating speed** (`operateSpeed`): ratio of travel distance to travel time, DRIVE_MILE/DRIVE_TIME.
- **Manhattan distance** (`Manhattan`): sum of absolute longitude and latitude differences between origin and destination, serving as a proxy for spatial displacement.
- **Administrative mapping** (`StartCode` / `EndCode`): origin and destination mapped from grids to corresponding administrative district codes.
- **Abnormal order label** (`IsAbnormal` or `IS_ANOMALY_5PCT`): binary indicator where 1 represents an abnormal or detour order and 0 represents a normal trip. In this work, a subset of experiments uses a label derived from extreme value theory and a fixed top-5% anomaly ratio.

3.2 Feature Summary

The main features used for modeling are summarized in Table 2.

4 Experimental Results

4.1 Problem Formulation

We formulate taxi order anomaly detection as a binary classification task. For each completed trip, let $\mathbf{x} \in \mathbb{R}^d$ denote the feature vector constructed from trip-level order data, including origin and destination coordinates, trip distance, travel time, fare, and temporal indicators. Let $y \in \{0, 1\}$ be the corresponding label, where $y = 1$ indicates an anomalous trip such as a detour or an overcharged trip, and $y = 0$ indicates a normal trip.

Given a labeled dataset

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N},$$

the goal is to learn a classifier

$$f_\theta : \mathbb{R}^d \to \{0, 1\},$$

Table 2: Summary of main features used in the modeling framework.

Feature Name	**Description**	**Unit**	**Example**
DEP_LON	Longitude of the trip origin	degrees	118.702884
DEP_LAT	Latitude of the trip origin	degrees	32.117048
DEST_LON	Longitude of the trip destination	degrees	118.727194
DEST_LAT	Latitude of the trip destination	degrees	32.112422
DRIVE_MILE	Distance of the trip, measured along the actual driving path	km	3.01
PRICE	Total fare paid for the trip	CNY	9
DRIVE_TIME	Travel time from origin to destination, computed from timestamps	minutes	5.82
IsPeakTime	Whether the order occurred during peak commuting hours	0/1	1
isWeekend	Indicator of weekend trips (1 = weekend, 0 = weekday)	0/1	1
isTime	Indicator of early-morning trips (1 = before 8 a.m., 0 = otherwise)	0/1	1
isLong	Trip-length category: 0 = $\leq$3 km, 1 = 3–12 km, 2 = 12–24 km, 3 = >24 km	categorical	1
operateSpeed	Average operating speed, DRIVE_MILE/DRIVE_TIME	km/min	0.52
Manhattan	Manhattan distance between origin and destination	degrees	0.024
IsAbnormal	Label indicating whether the trip is abnormal	0/1	1

that predicts the anomaly label for each trip,

$$\hat{y}_i = f_\theta(\mathbf{x}_i).$$

4.2 Baseline Models and Hyperparameters

Three baseline models are considered in this study:

- **Decision Tree**: a tree-based classifier that uses Gini impurity as the split criterion. The maximum depth and the minimum number of samples per node are constrained in order to limit model complexity and reduce overfitting.
- **Support Vector Machine (SVM)**: a classifier with a radial basis function kernel. Class weights are adjusted to account for label imbalance, and probability outputs are enabled so that ROC-AUC and PR-AUC can be computed in a consistent way.
- **Isolation Forest**: an ensemble of isolation trees that is commonly used for anomaly detection. In this work it is configured to produce anomaly scores and corresponding binary labels based on a specified contamination rate.

For each model, key hyperparameters and search ranges are summarized in Table 3.

4.3 Evaluation Metrics

Anomalies are treated as the positive class. Let TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively. We report the following metrics:

- **Accuracy.**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

  Accuracy measures the overall fraction of correctly classified trips. In heavily imbalanced anomaly detection problems, however, a trivial classifier that always predicts normal can achieve high accuracy while completely failing to detect anomalies, so accuracy alone is not informative.

- **Precision.**

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

  Precision quantifies how many of the trips flagged as anomalous are truly anomalous. High precision means that manual inspection is focused on genuinely suspicious trips and the number of false alarms is limited.

- **Recall.**

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

  Recall measures the fraction of true anomalies that are successfully detected. Missing anomalies, that is a large FN, can be much more costly than raising additional alarms, so recall is often a primary objective.

Table 3: Hyperparameter search space for the three anomaly detection models

Model	Hyperparameter	Search space	Motivation
Decision Tree	`max_depth`	{`None`, 5, 10}	Controls model complexity and the bias–variance trade-off; limiting depth reduces overfitting.
	`min_samples_leaf`	{1, 5, 10}	Avoids very small leaves and stabilizes decisions on rare patterns.
	`min_samples_split`	{2, 10, 20}	Prevents splitting on very small nodes and improves robustness to noise.
	`class_weight`	`balanced`	Increases the influence of the minority anomaly class during training.
SVM	`kernel`	`rbf`	Allows flexible non-linear decision boundaries in feature space.
	C	{0.1, 1.0, 10.0}	Balances margin smoothness against fitting the training data.
	γ	{`scale`, `auto`}	Controls the effective width of the RBF kernel around each sample.
	`class_weight`	`balanced`	Mitigates label imbalance by penalizing errors on anomalies more strongly.
	`probability`	`True`	Provides continuous scores used for ROC-AUC and PR-AUC evaluation.
Isolation Forest	`n_estimators`	{100, 200}	More trees yield more stable anomaly scores at additional computational cost.
	`max_samples`	{0.5, 1.0}	Subsampling per tree decorrelates trees and improves isolation of rare anomalies.
	`contamination`	{0.03, 0.05, 0.08}	Sets the expected anomaly proportion and calibrates the score threshold.

- **F1-score.**
$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$
The F1-score is the harmonic mean of precision and recall and provides a single summary of their trade-off at a given decision threshold.

- **ROC-AUC.** The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ as the decision threshold is varied. ROC-AUC summarizes the area under this curve and measures the model's ranking ability over all possible thresholds. However, in highly imbalanced settings, ROC-AUC may appear optimistic, because the false positive rate can remain small even when the absolute number of false alarms is large.

- **PR-AUC.** The precision–recall (PR) curve plots precision versus recall as the decision threshold varies, and PR-AUC is the area under this curve. Unlike ROC-AUC, PR-AUC focuses directly on the performance for the positive, anomalous class and is particularly informative in rare-event detection. A higher PR-AUC indicates that the model is able to rank truly anomalous trips near the top of its anomaly score distribution, while maintaining a reasonable precision level as recall increases.

Overall, while accuracy and ROC-AUC provide a global view of classification performance, precision, recall, F1-score, and especially PR-AUC are more relevant for anomaly detection, because they explicitly characterize the trade-off between missed anomalies and false alarms in the highly imbalanced setting. In the subsequent hyperparameter tuning and model comparison, PR-AUC is therefore adopted as the primary selection criterion, and the remaining metrics are reported as complementary indicators of overall classification quality.

4.4 Hyperparameter Tuning Strategy

Hyperparameter tuning is conducted after fixing the feature representation and the data preprocessing pipeline, and before reporting any final performance. The overall procedure is as follows:

1. Fix the input feature set and preprocessing scheme. Continuous variables are standardized to zero mean and unit variance, while discrete variables are converted to one-hot vectors.

2. Apply stratified K-fold cross-validation with $K = 5$ to the full labeled dataset so that each fold preserves the anomaly proportion. For each model family and for each candidate hyperparameter configuration in Table 3, the model is trained on four folds and evaluated on the remaining fold, and this procedure is repeated until every fold has served once as the validation fold.

3. For each configuration, compute the precision–recall area under the curve, PR-AUC, on the anomaly class for each validation fold and then average these values across the five folds. The configuration with the highest mean PR-AUC is selected as the best setting for that model family.

4. With the selected hyperparameters, summarize performance by averaging the fold-wise metrics, including accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC, over the five validation folds. These averaged values are reported as the final cross-validated performance of each model family.

The choice of PR-AUC as the primary tuning criterion is motivated by the strong class imbalance in the anomaly detection setting. Anomalous trips form only a small fraction of all trips, so traditional metrics such as accuracy or even ROC-AUC may appear high even when many anomalies are missed. PR-AUC focuses directly on the precision and recall of the positive class and therefore better reflects how well a model can rank and detect rare anomalous trips.

After hyperparameters have been tuned separately for each model family using the 5-fold cross-validation procedure, the resulting tuned models are compared based on their cross-validated performance. The final detector is chosen as the model that achieves the most favorable anomaly-oriented performance, with particular emphasis on PR-AUC, recall, and F1-score for the anomaly class, and can then be refitted on the entire labeled dataset for visualization and downstream analysis.

4.5 Overall Model Comparison

Based on the grid search described above, the selected hyperparameters for each model family are:

- **Decision Tree**: `max_depth = 10`, `min_samples_leaf = 10`, `min_samples_split = 2`.
- **SVM**: `C = 10.0`, `gamma = auto`.
- **Isolation Forest**: `contamination = 0.03`, `max_samples = 0.5`, `n_estimators = 200`.

These settings are used to produce the cross validation results reported in Table 4 and form the basis for the subsequent comparison and analysis. The Table 4 compares the three models under 5-fold cross-validation on the labeled dataset, where anomalies correspond to the empirically selected top-5% tail based on residual-based extreme value analysis. From Table 4, several patterns emerge:

Table 4: Performance of models

Metric	Decision Tree	SVM	Isolation Forest
Accuracy	**0.960** $\pm$ 0.002	0.925 $\pm$ 0.005	0.544 $\pm$ 0.003
Precision	**0.669** $\pm$ 0.029	0.400 $\pm$ 0.019	0.096 $\pm$ 0.003
Recall	0.496 $\pm$ 0.053	0.864 $\pm$ 0.025	**0.916** $\pm$ 0.028
F1-score	**0.568** $\pm$ 0.036	0.547 $\pm$ 0.019	0.174 $\pm$ 0.005
ROC-AUC	0.921 $\pm$ 0.015	**0.957** $\pm$ 0.007	0.831 $\pm$ 0.025
PR-AUC	0.596 $\pm$ 0.043	**0.679** $\pm$ 0.012	0.316 $\pm$ 0.051

Notes: values are mean $\pm$ standard deviation over the 5 folds.

From Table 4, several observations can be made:

- **Decision Tree** attains the highest accuracy (0.960 ± 0.002), precision (0.669 ± 0.029), and F1-score (0.568 ± 0.036) among the three models. This shows that most of the trips it flags as anomalous are indeed labeled as anomalies, and overall classifications are very often correct. However, its recall on anomalies is only moderate (0.496 ± 0.053), which means that a nontrivial fraction of anomalous trips are still missed. Its ROC-AUC and PR-AUC (0.921 ± 0.015 and 0.596 ± 0.043) are respectable but not the best.
- **SVM** provides the most balanced and anomaly-oriented performance. It achieves high accuracy (0.925 ± 0.005) and precision (0.400 ± 0.019), together with very strong recall (0.864 ± 0.025). As a result, its F1-score (0.547 ± 0.019) is only slightly lower than that

of the Decision Tree, while its ROC-AUC and PR-AUC are the highest (0.957 ± 0.007 and 0.679 ± 0.012). These results indicate that the SVM ranks anomalous trips most effectively and offers the best trade-off between detecting anomalies and avoiding false positives.

- **Isolation Forest** achieves the highest recall on anomalies (0.916 ± 0.028), meaning that it successfully identifies many of the labeled anomalous trips. However, its accuracy, precision, F1-score, ROC-AUC, and PR-AUC are all substantially lower than those of the other two models (for example, precision 0.096 ± 0.003 and PR-AUC 0.316 ± 0.051), which implies a large number of false alarms and weak overall discrimination.

Overall, the SVM offers the most favorable compromise between anomaly coverage and false-alarm rate, and is therefore selected as the final model for downstream visualization and analysis.

4.6 Feature Ablation Study

To understand the contribution of different features to the SVM performance, a series of feature ablation experiments are conducted. The base feature set used in the SVM anomaly detector is drawn from Table 2 and includes both spatial variables (`DEP_LON`, `DEP_LAT`, `DEST_LON`, `DEST_LAT`, `Manhattan`), trip-level distance and time variables (`DRIVE_MILE`, `DRIVE_TIME`), cost information (`PRICE`), and several contextual or behavioral indicators (such as `hr`, `isTime`, `isLong`, `IsPeakTime`, `isWeekend`).

These ablations target three main groups of information: (i) spatial structure (coordinates, Manhattan distance, and administrative codes), (ii) temporal and behavioral context (hour-of-day, peak-time, weekend, and trip-length indicators), and (iii) cost-related factors (distance, time, fare, and derived operating speed). By removing one component at a time, we can quantify its marginal contribution to the anomaly detection performance of the SVM. The corresponding 5-fold cross-validation results are reported in Table 5, from which several insights emerge:

- **Temporal context is highly informative.** When temporal indicators such as `IsPeakTime`, `isWeekend`, `isTime` and `isLong` are removed, the values of PR AUC and F1 score show a consistent decrease or at best only very slight improvement. For example, removing `isWeekend` or `isTime` yields PR AUC values close to that of the full model but does not improve F1, which means that temporal structure still contributes useful information for distinguishing normal and abnormal trips.

- **Distance, time and price form the core signal.** Removing `DRIVE_MILE`, `DRIVE_TIME` or `PRICE` causes clear drops in both F1 score and PR AUC compared with the full feature set. This shows that anomalies are mainly reflected in the joint pattern of distance, travel time and fare rather than in any single quantity. Once any element of this triplet is removed, the detector becomes less effective at capturing detours and overcharging.

- **Spatial information and speed are complementary.** Eliminating origin and destination coordinates or the Manhattan distance leads to only moderate degradation in PR AUC, and the SVM still keeps competitive performance. Similarly, removing `operateSpeed` slightly changes F1 score and PR AUC but does not alter the overall ranking of feature importance. These results suggest that spatial location and average speed refine the decision boundary and help explain regional differences, while the dominant discriminative power still comes from the distance–time–price relationship together with temporal indicators.

Table 5: Feature ablation results for the SVM anomaly detector

Feature removed	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
DEP/DEST	0.902 ±0.001	0.340 ±0.004	**0.913 ±0.033**	0.495 ±0.008	0.957 ±0.014	0.639 ±0.040
DRIVE_MILE	0.924 ±0.006	0.400 ±0.020	0.862 ±0.021	0.546 ±0.019	0.956 ±0.007	0.678 ±0.012
PRICE	0.886 ±0.004	0.292 ±0.006	0.819 ±0.026	0.430 ±0.006	0.920 ±0.006	0.527 ±0.016
DRIVE_TIME	0.922 ±0.006	0.391 ±0.017	**0.870 ±0.019**	0.539 ±0.015	0.955 ±0.006	0.679 ±0.016
IsPeakTime	0.925 ±0.004	**0.403 ±0.015**	**0.870 ±0.018**	**0.550 ±0.013**	**0.958 ±0.009**	**0.688 ±0.018**
isWeekend	**0.926 ±0.005**	**0.407 ±0.020**	0.858 ±0.028	**0.551 ±0.020**	0.956 ±0.007	0.677 ±0.012
isTime	**0.928 ±0.005**	**0.412 ±0.021**	0.850 ±0.026	**0.555 ±0.021**	0.955 ±0.007	0.674 ±0.014
isLong	0.924 ±0.004	0.395 ±0.015	0.841 ±0.023	0.538 ±0.017	0.952 ±0.008	0.631 ±0.014
operateSpeed	0.925 ±0.004	**0.402 ±0.016**	**0.874 ±0.015**	**0.551 ±0.016**	0.957 ±0.008	**0.690 ±0.012**
Manhattan	0.919 ±0.007	0.376 ±0.024	0.820 ±0.053	0.515 ±0.029	0.943 ±0.011	0.631 ±0.032
keep all features	0.925 ±0.005	0.400 ±0.019	0.864 ±0.025	0.547 ±0.019	0.957 ±0.007	0.679 ±0.012

Notes: values are mean ± standard deviation over the 5 folds.
Red numbers indicate performance higher than the "keep all features" baseline.

Overall, the ablation study confirms that the SVM anomaly detector relies on a combination of feature groups. Spatial features such as coordinates and Manhattan distance, temporal and behavioural features such as peak hour, weekend and trip length, and cost related features such as distance, time and fare jointly support accurate detection. No single group is sufficient on its own, and it is the interaction among these components that enables reliable identification of abnormal taxi orders.

Table 6: Effect of jointly removing `operateSpeed` and `IsPeakTime` on SVM performance.

Setting	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
keep all features	0.925	0.400	0.864	0.547	0.957	0.679
remove two features	**0.926**	**0.406**	**0.878**	**0.555**	**0.959**	**0.696**
Difference	+0.001	+0.006	+0.014	+0.008	+0.002	+0.017

Table 6 compares the baseline SVM using the full feature set with a variant where `operateSpeed` and `IsPeakTime` are removed simultaneously. All six metrics improve slightly when these two features are dropped. In particular, recall increases from 0.864 to 0.878 and PR-AUC increases from 0.679 to 0.696. This suggests that `operateSpeed` and `IsPeakTime` do not provide additional useful information beyond the remaining features and may even introduce noise or redundancy into the classifier.

4.7 Evaluation of model performance on the full dataset

The numerical cross validation results in Table 4, together with the feature ablation study in Table 5, show that the SVM offers the most balanced trade off between recall, precision, F1 score and PR AUC when using the reduced feature set that excludes `operateSpeed` and `IsPeakTime`. Based on these findings, we adopt this reduced feature configuration as the final detector, retrain the SVM with the selected hyperparameters on the full labeled dataset, and then present a series of graphical summaries that illustrate how the three models behave across different operating regimes and how the final SVM performs in detail on the full dataset.

4.7.1 Precision–recall and ROC characteristics

Figures 2 and 3 report the precision–recall curve and the ROC curve of the three models when trained on the full labeled dataset.

In Figure 2 the curves of the SVM and the decision tree lie far above the random baseline across almost the entire recall range, which shows that both models can rank anomalous trips near the top of the score spectrum. In the low recall region where only a small fraction of trips is flagged, the decision tree achieves very high precision, which means that the most suspicious trips identified by the tree are almost all true anomalies. As recall grows, the precision of the tree drops more quickly, whereas the SVM maintains noticeably higher precision at medium and high recall. The curve of Isolation Forest stays well below the other two curves, which indicates a weaker ability to separate anomalous and normal trips when the operating threshold varies.

Figure 3 shows that the SVM and the decision tree both achieve very high ROC AUC values close to one. Their curves concentrate near the upper left corner, which means that they can reach high true positive rates at modest false positive rates. The ROC curve of Isolation Forest remains farther from the ideal corner and its AUC is much smaller, around 0.80. Although ROC AUC is

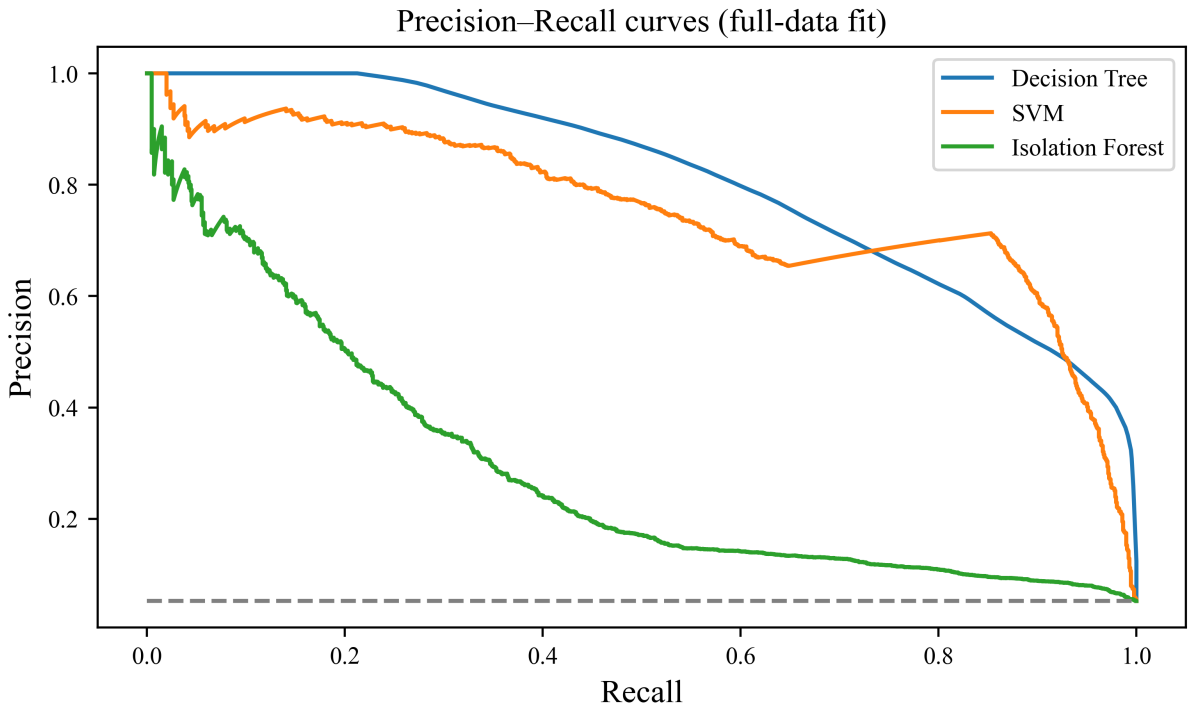

Figure 2: Precision–recall curves of the three models fitted on the full dataset

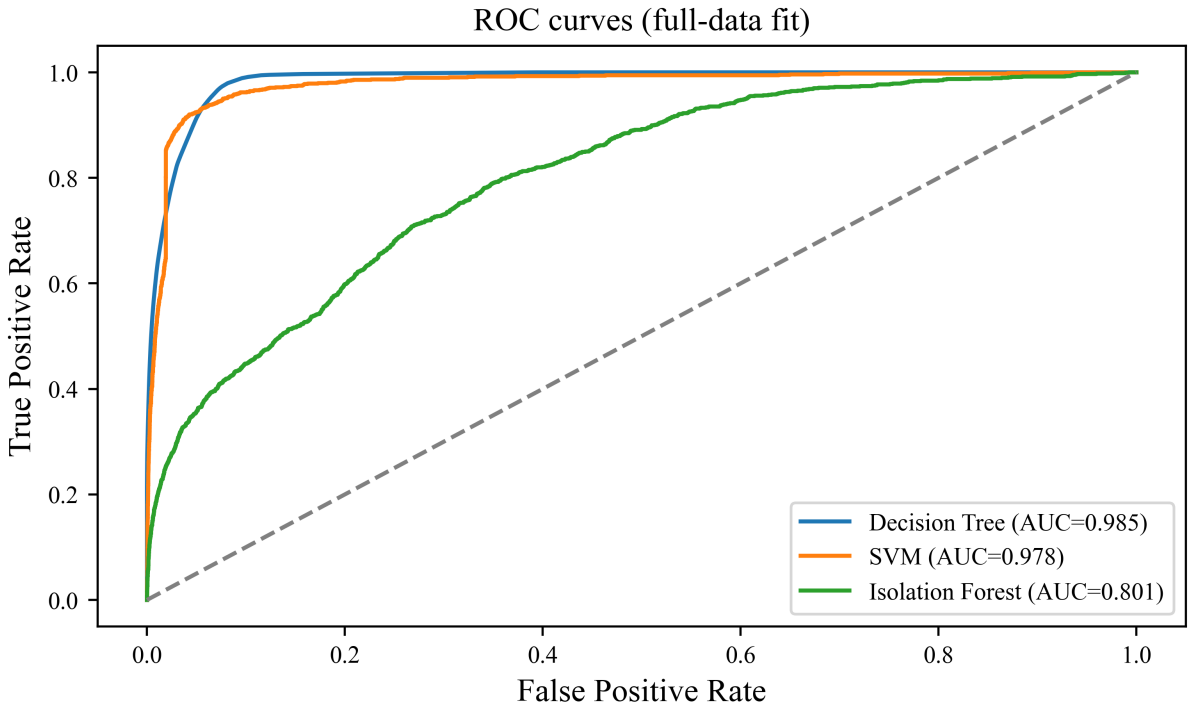

Figure 3: ROC curves of the three models fitted on the full dataset

less sensitive to class imbalance than PR AUC, these curves again confirm that the SVM and the decision tree have much stronger ranking ability than Isolation Forest.

4.7.2 Confusion matrix and score distribution of the SVM

To inspect the concrete classification behaviour of the chosen model, Figures 4 and 5 summarize the SVM performance at a fixed operating threshold on the full dataset.

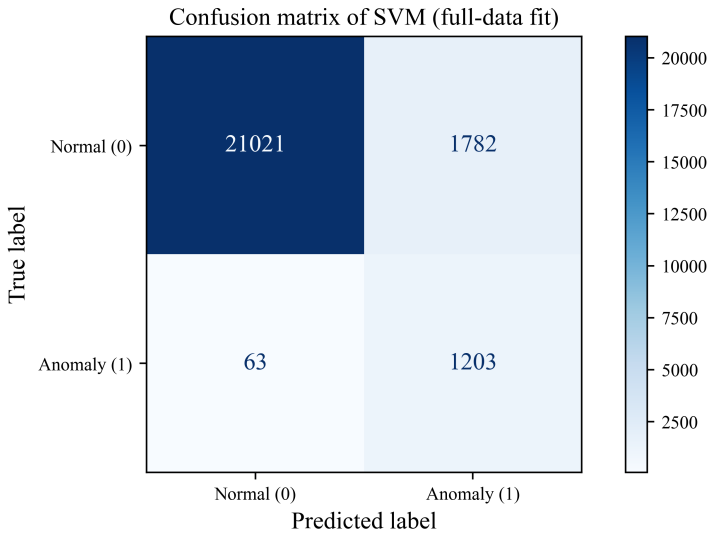

Figure 4: Confusion matrix of the SVM on the full dataset at the chosen operating threshold

The confusion matrix in Figure 4 reveals that the SVM successfully identifies 1249 anomalous trips and misses only 53, which corresponds to a very high recall on the anomaly class. At the same time 1883 normal trips are incorrectly flagged as anomalous, which yields a moderate precision around 0.40. This operating point is therefore tuned toward high anomaly coverage with an acceptable number of false alarms, which is appropriate for safety and fraud detection applications where missing anomalies is more costly than further manual inspection.

Figure 5 plots the histogram of SVM anomaly scores for normal trips and anomalous trips. Most normal trips receive scores very close to zero and form a sharp peak near the origin. Anomalous trips are shifted toward larger scores and include a subset with scores close to one. The two distributions therefore exhibit a clear separation with some overlap. This separation justifies the use of a single scalar threshold to obtain a high recall detector and motivates ranking based uses of the anomaly scores.

4.7.3 Effect of the decision threshold for the SVM

The impact of the decision threshold on SVM performance is illustrated in Figure 6.

In Figure 6, when the threshold is very small almost all trips are labeled as anomalous. Recall is close to one, but precision is low since many normal trips are flagged. As the threshold increases, precision improves steadily while recall decreases. The F1 score rises from a low value, reaches a maximum near an intermediate threshold and then declines again for very high thresholds. The vertical line in the figure marks the threshold that maximizes the F1 score on the evaluation set. This curve highlights that the SVM detector can be tuned flexibly: one can move the operating

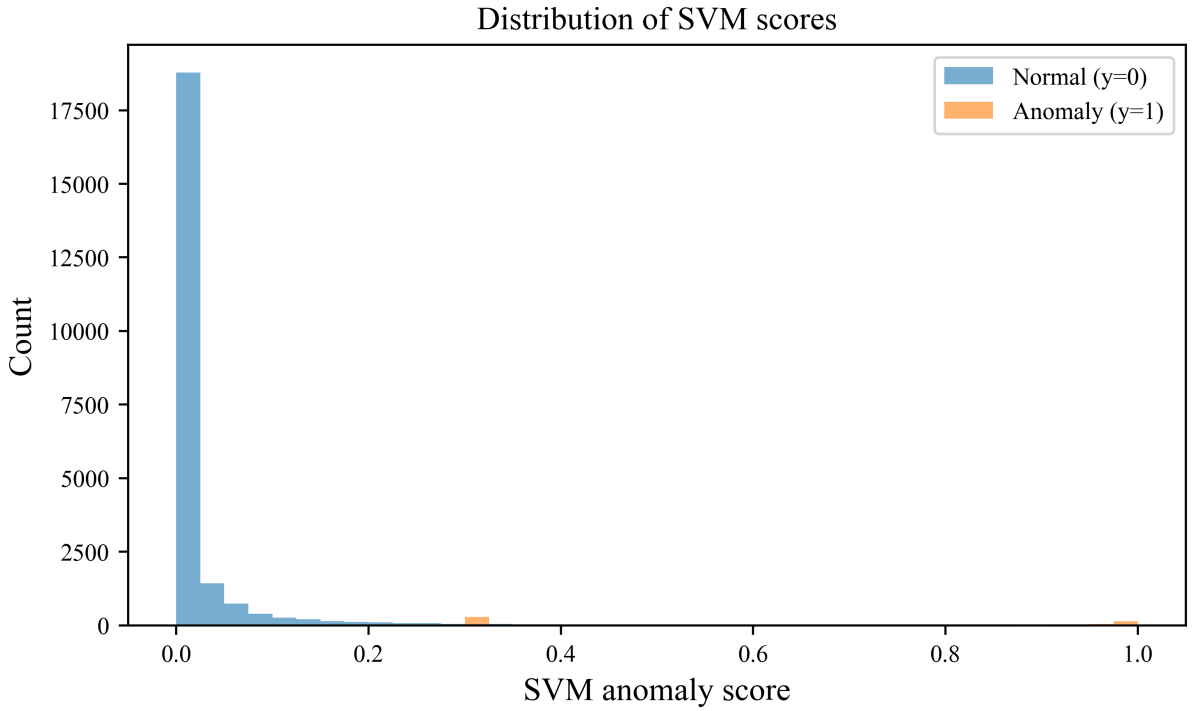

Figure 5: Distribution of SVM anomaly scores for normal trips and anomalous trips

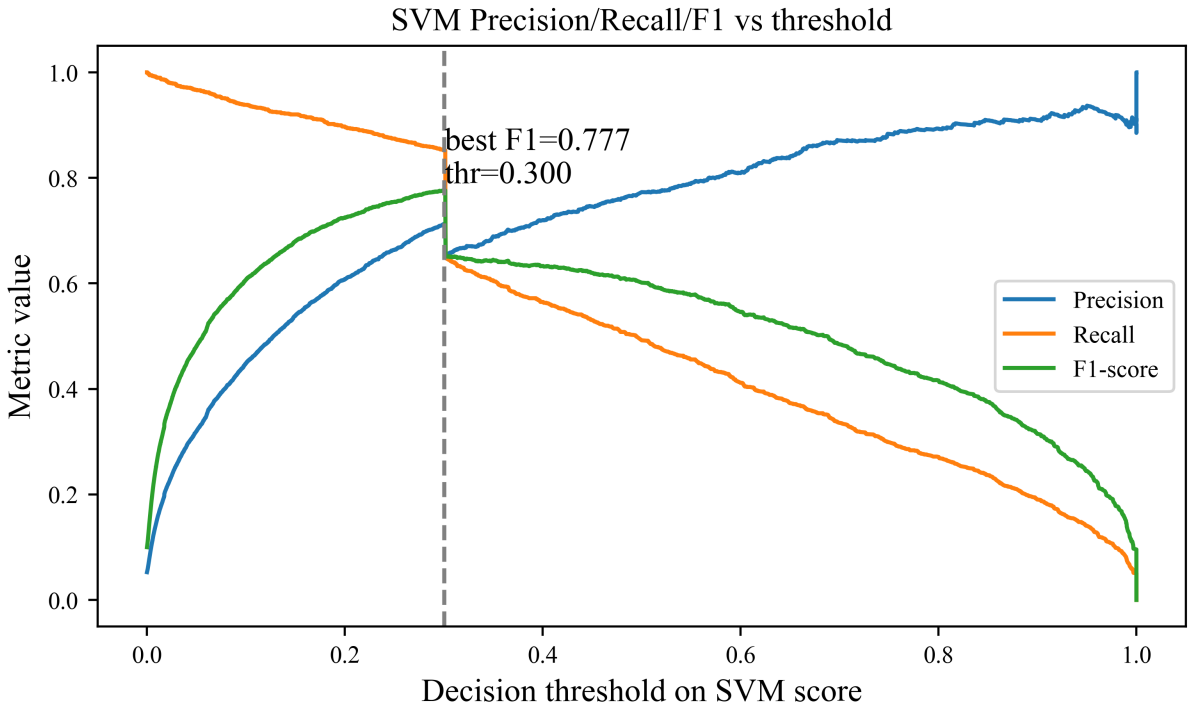

Figure 6: Precision, recall and F1 score of the SVM as functions of the decision threshold on the anomaly score

point toward higher recall for regulatory screening or toward higher precision for targeted inspection, depending on the available resources and tolerance for false alarms.

4.7.4 Top K inspection scenario

In many practical applications only a limited number of trips with the highest anomaly scores can be inspected by human experts. To assess this use case, Figures 7 and 8 compare the SVM and Isolation Forest under a top K inspection strategy.

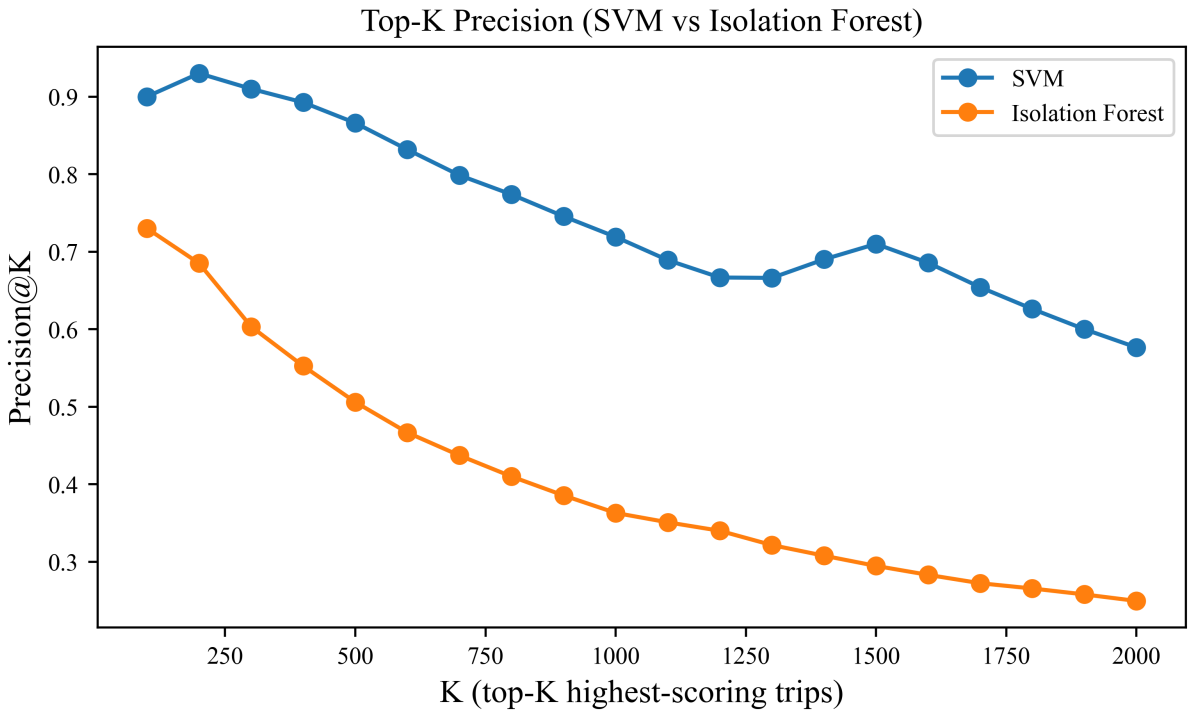

Figure 7: Top K precision of SVM and Isolation Forest when ranking trips by anomaly score

In Figure 7 the horizontal axis shows the number K of highest scoring trips that are selected for inspection and the vertical axis shows the precision among these K trips. For all values of K, the precision of the SVM is substantially higher than that of Isolation Forest. When K is between roughly 200 and 600, SVM precision remains near 0.9, which means that the vast majority of inspected trips are truly anomalous. As K grows, both curves decline because less extreme scores are included, yet SVM maintains a clear advantage across the entire range.

Figure 8 reports the corresponding recall at K, that is the fraction of all anomalous trips that fall into the top K ranked trips. The SVM consistently achieves a much higher recall at every inspection budget. For moderate values of K, it already recovers a large majority of anomalies, whereas Isolation Forest covers only a small fraction. Together with Figure 7, this result shows that under realistic inspection budgets the SVM can both concentrate true anomalies in the inspected set and cover a much larger share of them.

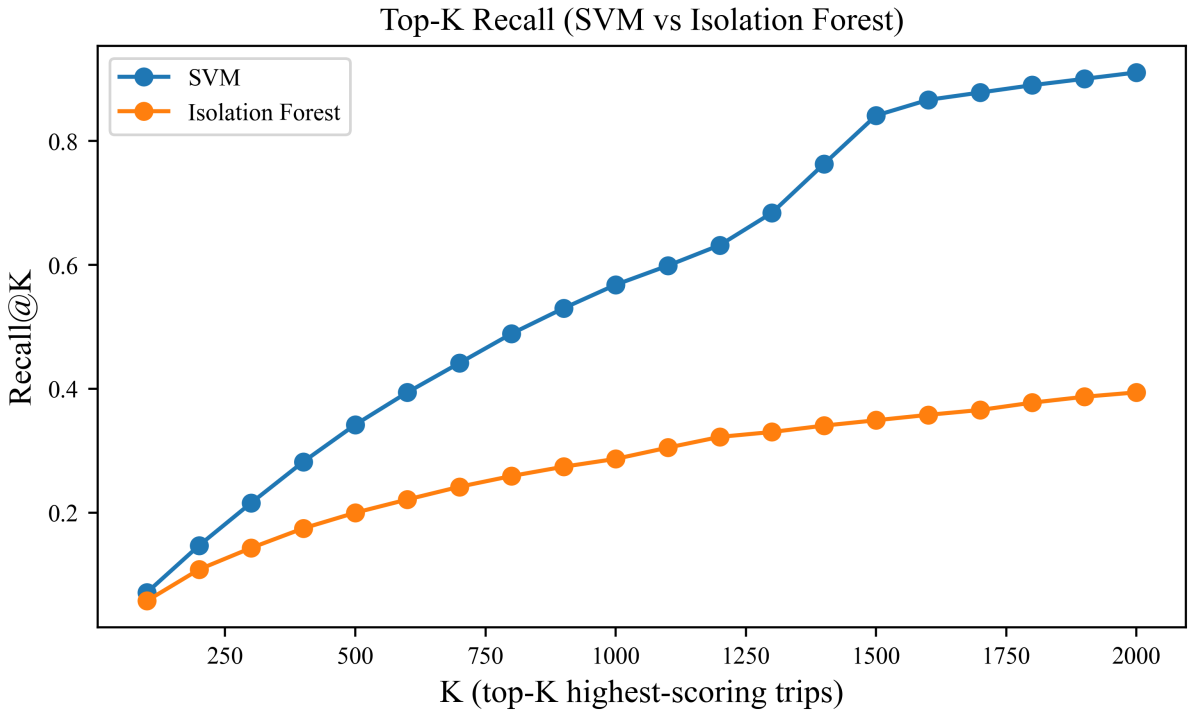

Figure 8: Top K recall of SVM and Isolation Forest when ranking trips by anomaly score

4.7.5 Summary of findings

The visual analyses in this subsection provide a detailed view of the models beyond aggregate cross validation scores. They show that the SVM combines strong ranking ability, flexible threshold tuning and excellent top K behaviour. The decision tree is highly competitive at very strict operating points that target only a handful of extreme anomalies, but loses precision as recall increases. Isolation Forest is clearly dominated by the supervised models in all considered metrics and scenarios. These observations validate the choice of SVM as the final detector for abnormal taxi orders and motivate the subsequent feature ablation study that focuses on understanding and improving this model.

5 Conclusion

This report presents a complete pipeline for taxi order anomaly detection using trip-level order data. After extensive data cleaning and feature engineering on a large-scale Nanjing ride-hailing dataset, multiple anomaly detection models were trained and evaluated. Among the baselines, the SVM with RBF kernel and class balancing offers the best trade-off between precision and recall, achieving the highest F1-score and ROC-AUC and competitive PR-AUC. Isolation Forest and Decision Tree provide useful complementary perspectives but are either too conservative (high precision, low recall) or too aggressive (high recall, low precision) for practical deployment.

The visual analysis of score distributions, confusion matrices, and Top-K recall underscores the usefulness of the SVM as a ranking model for generating prioritized inspection lists. Feature ablation experiments further highlight the importance of combining spatial, temporal, and cost-

related features.

6 Code availability

All scripts used for data preprocessing, model training, feature ablation, and result visualization are publicly available at `https://github.com/WangBo-2000/taxi-anomaly-detection.git`.