Regression Prediction of all Customers' Next Car-service Date and Odometer

Project Objective

To predict all customers' next car-service visit date and odometer reading, with the goal of limiting the deviation in predicted dates to within one month, thereby enabling the business dept to formulate data-driven marketing strategies.

Challenges

1. Low-frequency and low-loyalty user behavior, making patterns difficult to capture.
2. High volume of missing and noisy data, which significantly increases the complexity of achieving a prediction accuracy within a one-month deviation.
3. Designing an effective predictive model while preventing data leakage, ensuring both reliability and robustness.

Model Results

模型 LGB Model

ratio of predictions deviating within 30 days

规则全量 Statistics Rule Model

```
1 tongji1(dt,'天数偏差分类1')
```

天数偏差分类1	fenzi	fenmu	ratio
0 30-60天	68418	316906	0.215894
1 30天以内	126972	316906	0.400661
2 60-90天	40412	316906	0.127520
3 90天以上	81104	316906	0.255924

```
1 tongji1(rule,'天数偏差分类1')
```

天数偏差分类1	fenzi	fenmu	ratio
0 30-60天	99440	480498	0.206952
1 30天以内	146486	480498	0.304863
2 60-90天	63439	480498	0.132028
3 90天以上	171133	480498	0.356158

```
1 tongji1(dt,'公里数偏差分类1')
```

公里数偏差分类1	fenzi	fenmu	ratio
0 2500-5000公里	56457	316906	0.178151
1 2500公里以内	221241	316906	0.698128
2 5000-7500公里	15433	316906	0.048699
3 7500公里以上	23775	316906	0.075022

```
1 tongji2(rule,'公里数偏差分类1')
```

公里数偏差分类1	cnt	ratio
1 2500公里以内	412725	0.858953
0 2500-5000公里	35121	0.073093
3 7500公里以上	20658	0.042993
2 5000-7500公里	11994	0.024962

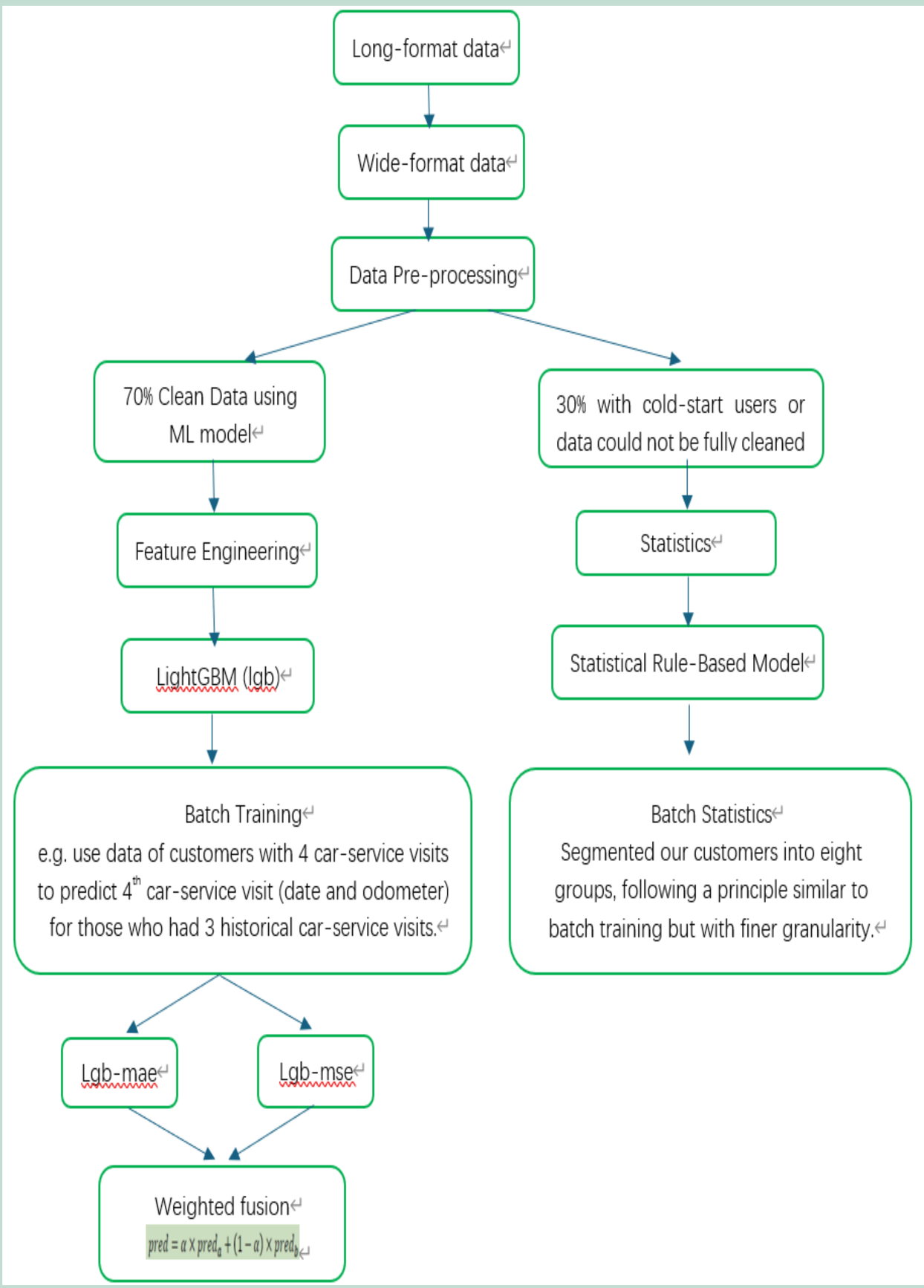
Summaries

1. Business logical understanding
2. wide-format structure data design
3. Creative batch-wise training and statistics
4. Innovative multi-metric weighted fusion for predictions
5. Deploying with PySpark in production enables faster processing and improved scalability.

Innovations

1. Transformed the original long-format dataset into a wide-format structure to prevent data leakage during training, avoiding inflated model performance.
2. Fed 70% of well-cleaned data into the algorithmic model, while the remaining 30%—including cold-start users or data that could not be fully cleaned—were processed using a statistical rule-based model.
3. Both algorithmic and statistical rule-based models innovatively employed granular, batch-wise training and statistical analysis.
4. Within the Machine Learning model, an innovative multi-metric weighted fusion approach was applied, enabling strong performance even with a single model.

Overall Model Framework



Alicee, Qin Lu

aliceeqin@gmail.com

+353 89 213 6352