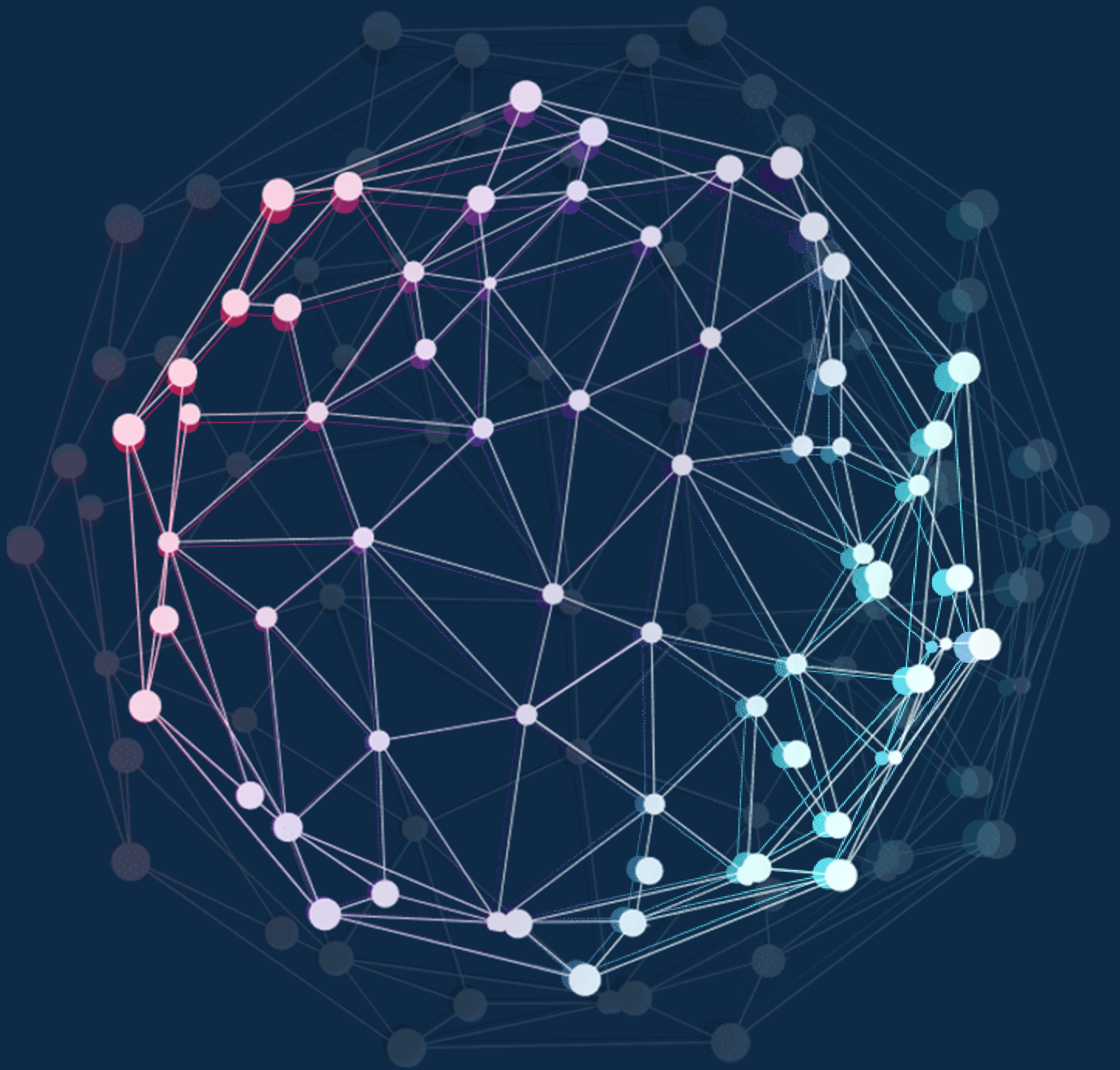
RIT Business Analytics Spring 2023 Competition



Presenter | Team SimonMiracle

Zhiwei Liang

Yuxiao Liao

Yitong Li

Xinyu Su

Tingting Liu

Problem Description

Scenario Description

Client:

RIT-Commerce

Task:

Address fake review problem and give advice

Fake reviews:

- Around 4% of online reviews
- Influence $152 billion transactions
- Generated by actual users

Detailed Task

- **Fake review detection**
- **Overall impact**
- **Practical implementations**

Data Preparation

Dataset

Dataset 1 ← Build Model

- Reviews Collection (2000-2020)
- 392,426 rows * 11 columns

Dataset 2 ← Predict "fake review"

- Reviews without the target (2000-2021)
- 43, 640 rows * 10 columns

Data Cleaning

Row training dataset 392426 rows × 11 columns

551 rows removed

Cleaned training dataset 391875 rows × 11 columns

Data Cleaning

Row training dataset

0	review_date	391875 non-null	datetime64[ns]
1	review_rating	391875 non-null	int64
2	number_of_photos	391875 non-null	int64
3	helpful_vote	391875 non-null	int64
4	reviewer_ID	391875 non-null	object
5	fake_asin	391875 non-null	int64
6	fake_review	391875 non-null	int64
7	product_ID	391875 non-null	int64
8	review_ID	391875 non-null	int64
9	review	391875 non-null	object

Data Preprocessing (cont.)

Feature Engineering:

$$\begin{tikzcd} & \textbf{Year} \\ \textbf{Date} \arrow[ur] \arrow[r] \arrow[dr] & \textbf{Month} \\ & \textbf{Day} \end{tikzcd}$$

Data Preprocessing (cont.)

Feature Engineering:

Text Variable

Gibberish & Punctuation

Stop-word

Word Stemming

Senti-Analysis

Data Preprocessing (cont.)

- Fake_asin
- Product_id

→ **Fake_asin*_Product_id (Product Feature)** → **Labeled_Product_idXXX**

labeled_product_id_82	...	labeled_product_id_427	labeled_product_id_428	labeled_product_id_429	labeled_product_id_436	labeled_product_id_438
0	...	0	0	0	0	0
0	...	0	0	0	0	0

Processed Training Dataset

[391875 rows x 52 columns]

Predictive model

Predictive Model Selection

Three models were constructed:

Support Vector Classifier (SVC)

- Highly effective for common text classification problems

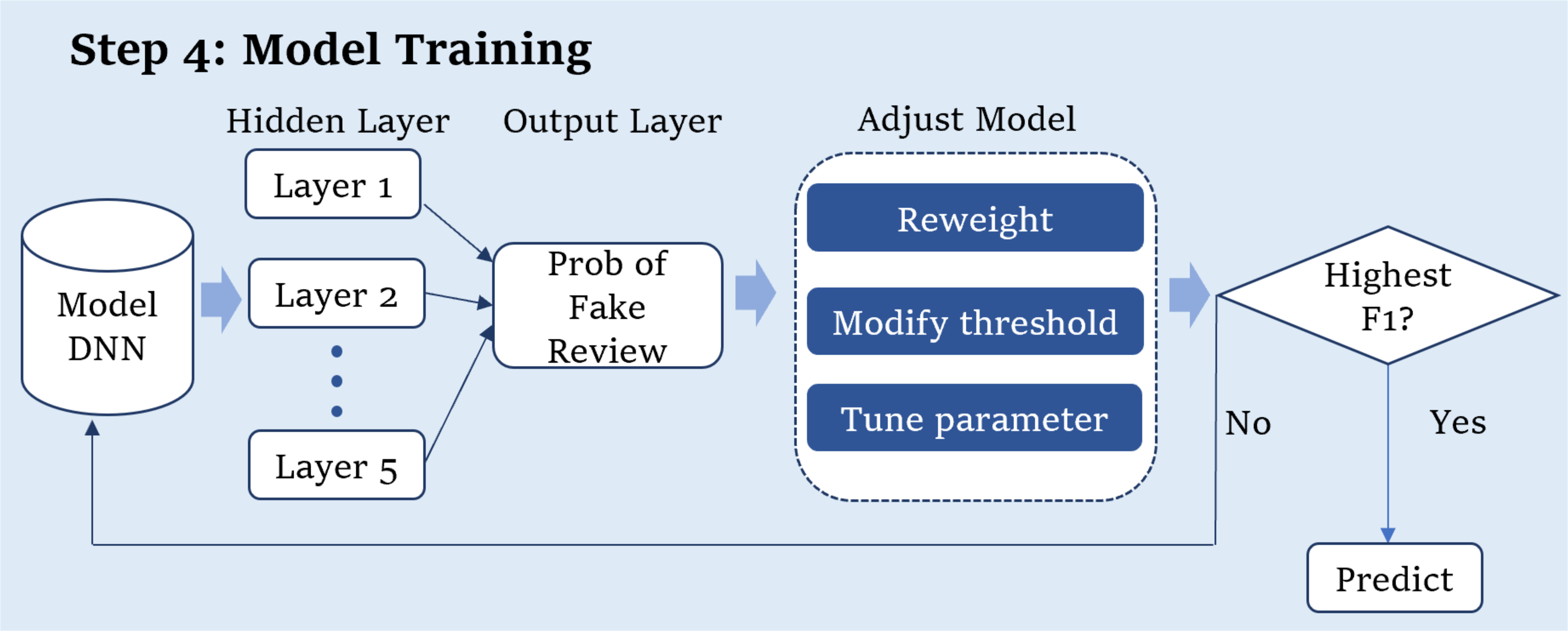
Recurrent Neural Network (RNN)

- Good at processing sequential data, especially natural language text

Deep Neural Network (DNN)

- Multiple layers
- Capable of modeling intricate relationships in large amounts of data

DNN Model Building Process

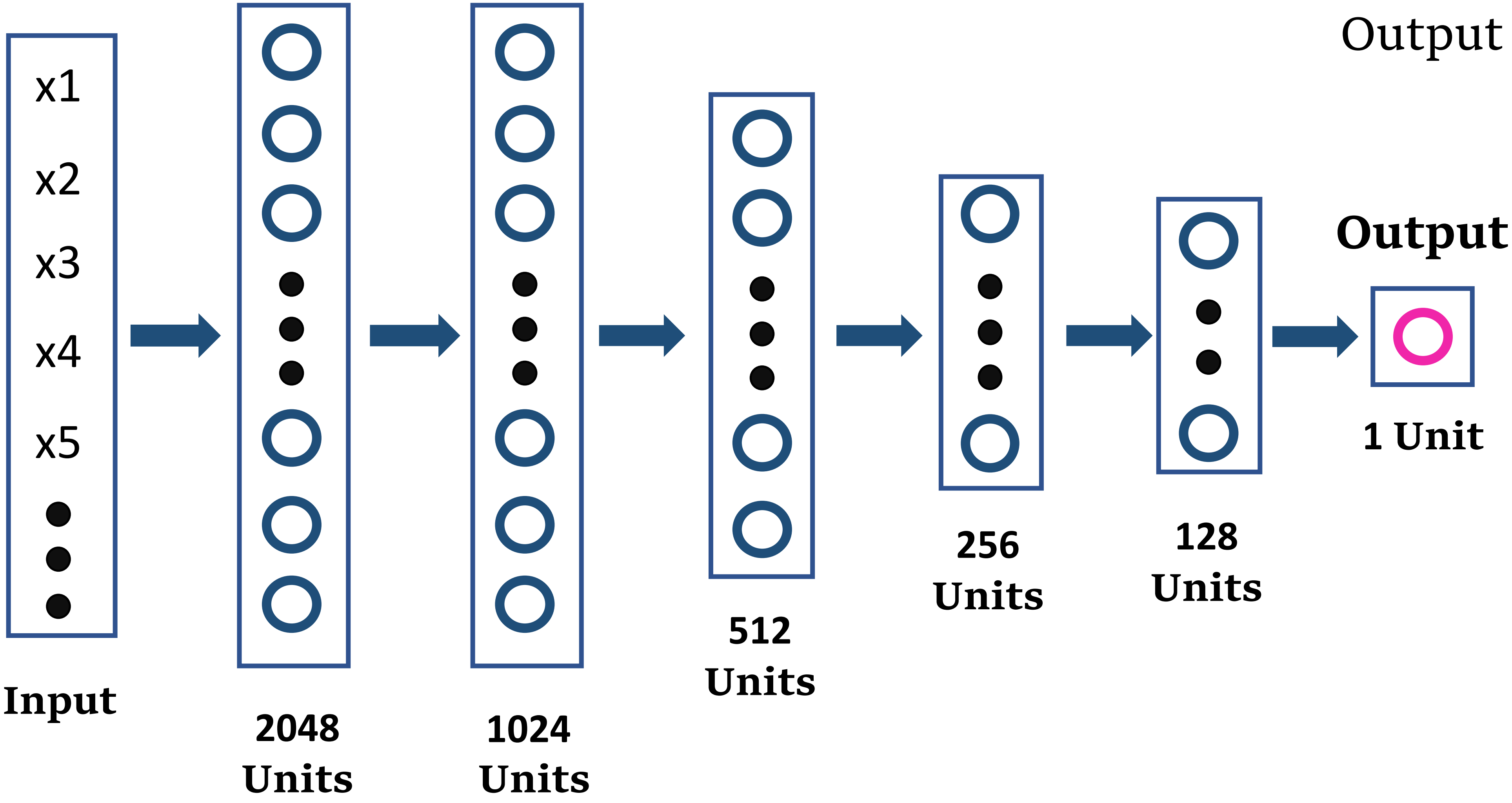


Layers and Units

Activation functions:

Hidden layers: **ReLu**

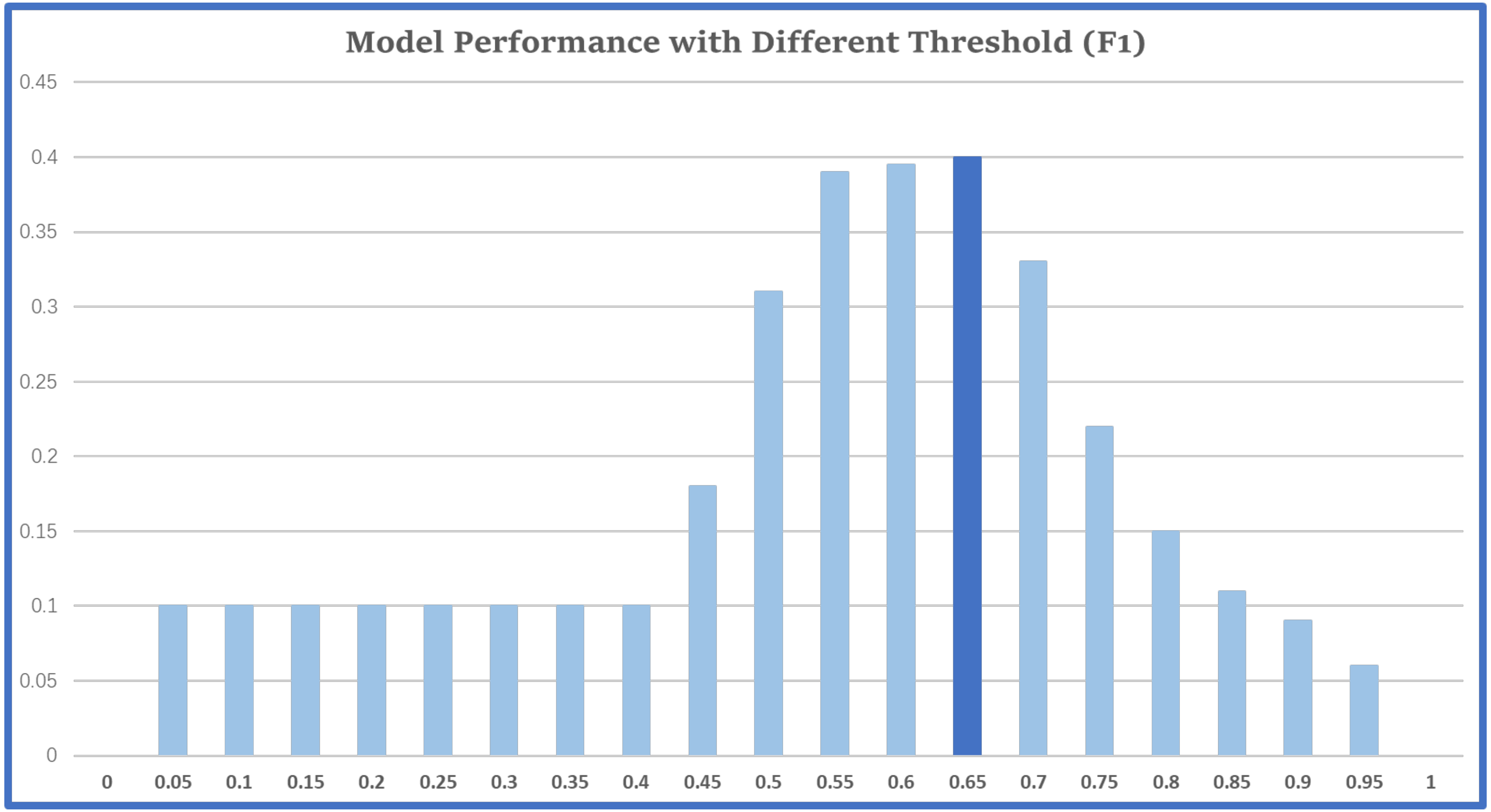
Output layer: **Sigmoid**



The probability that the review is fake

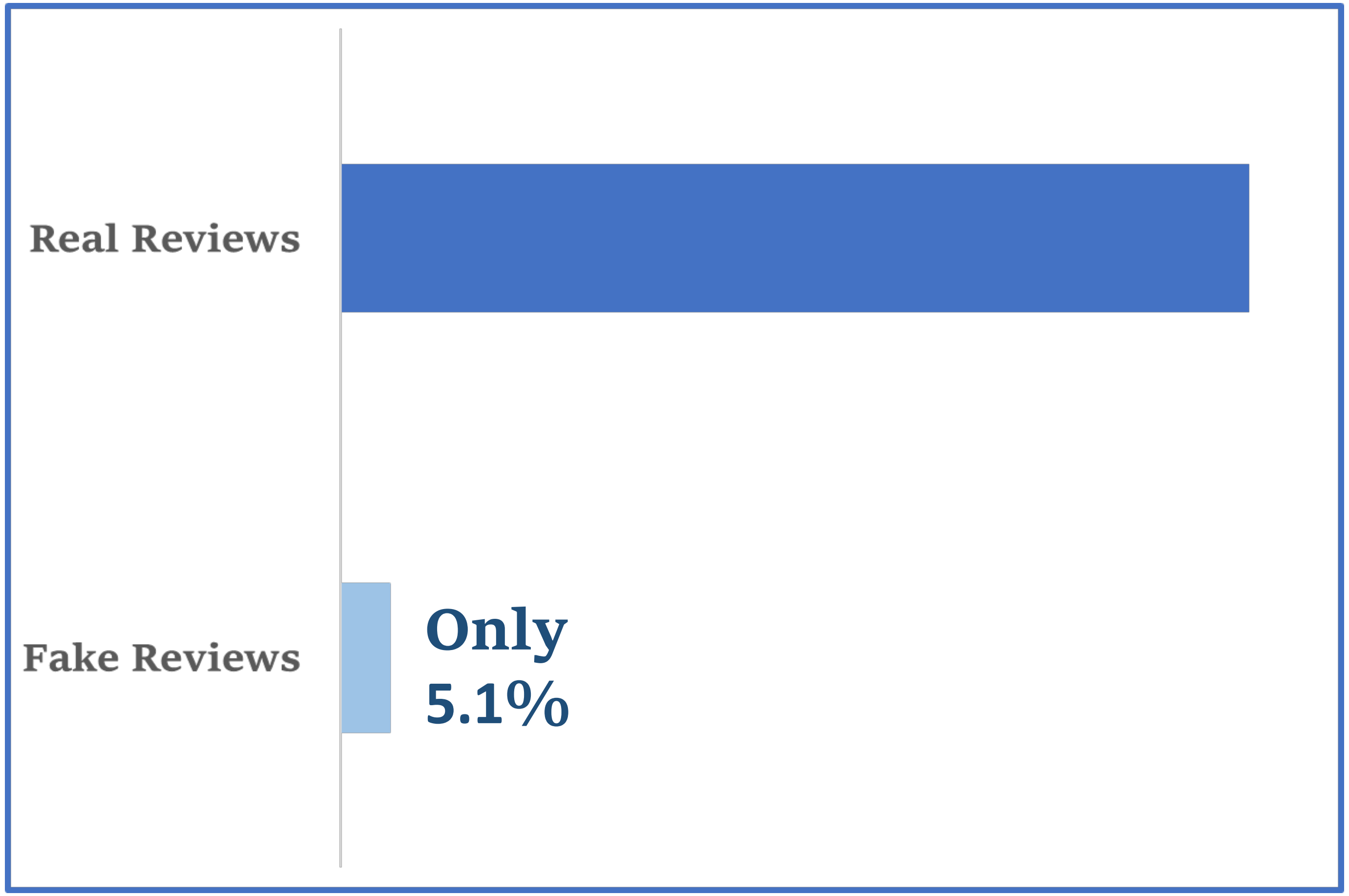
Adjusting Threshold

Output: the probability that the review is fake



Best Threshold = 0.65

Reweighting Unbalanced Dataset



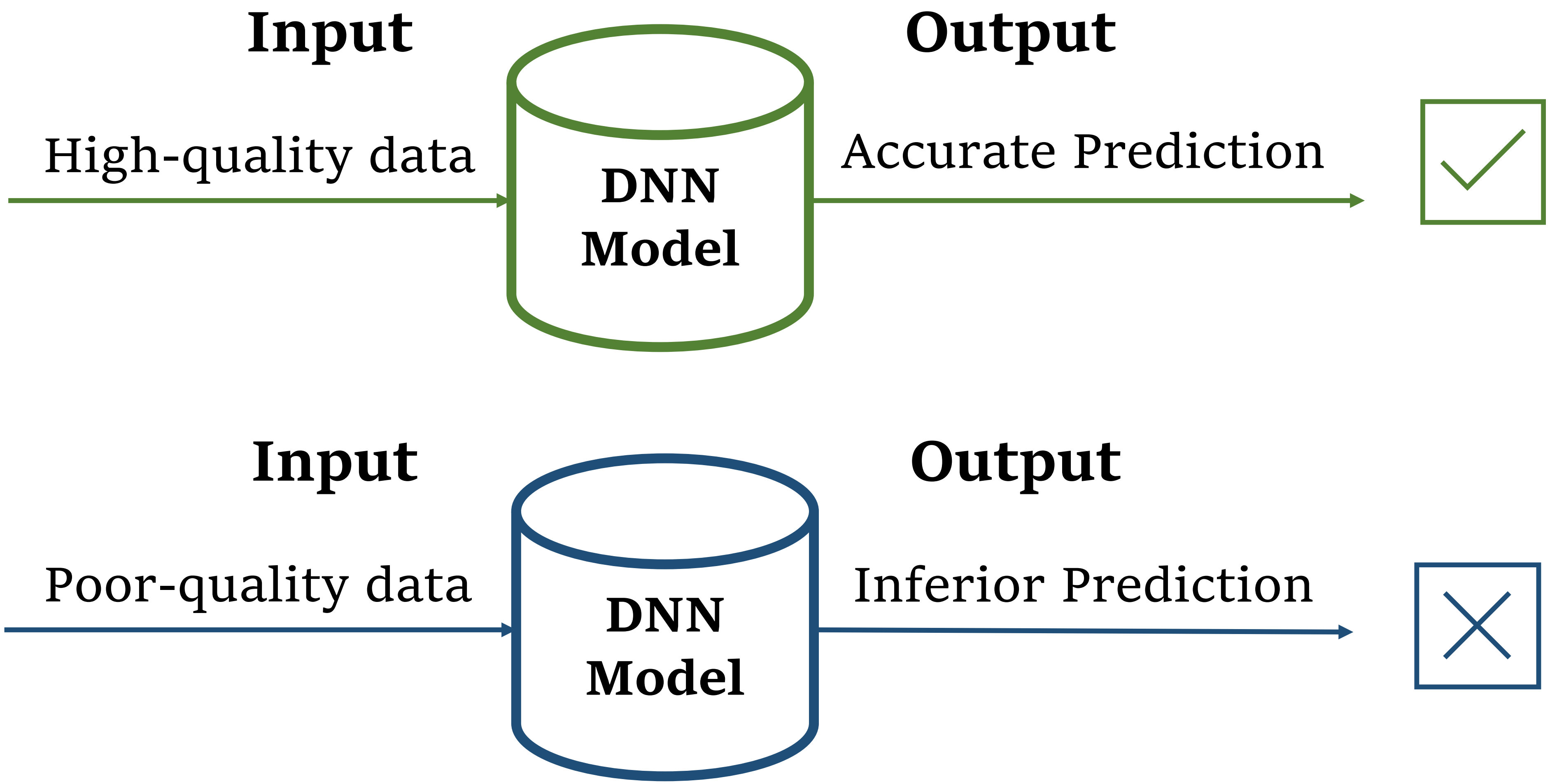
Before Reweighting

Private Score	Public Score
0.13103	0.13021

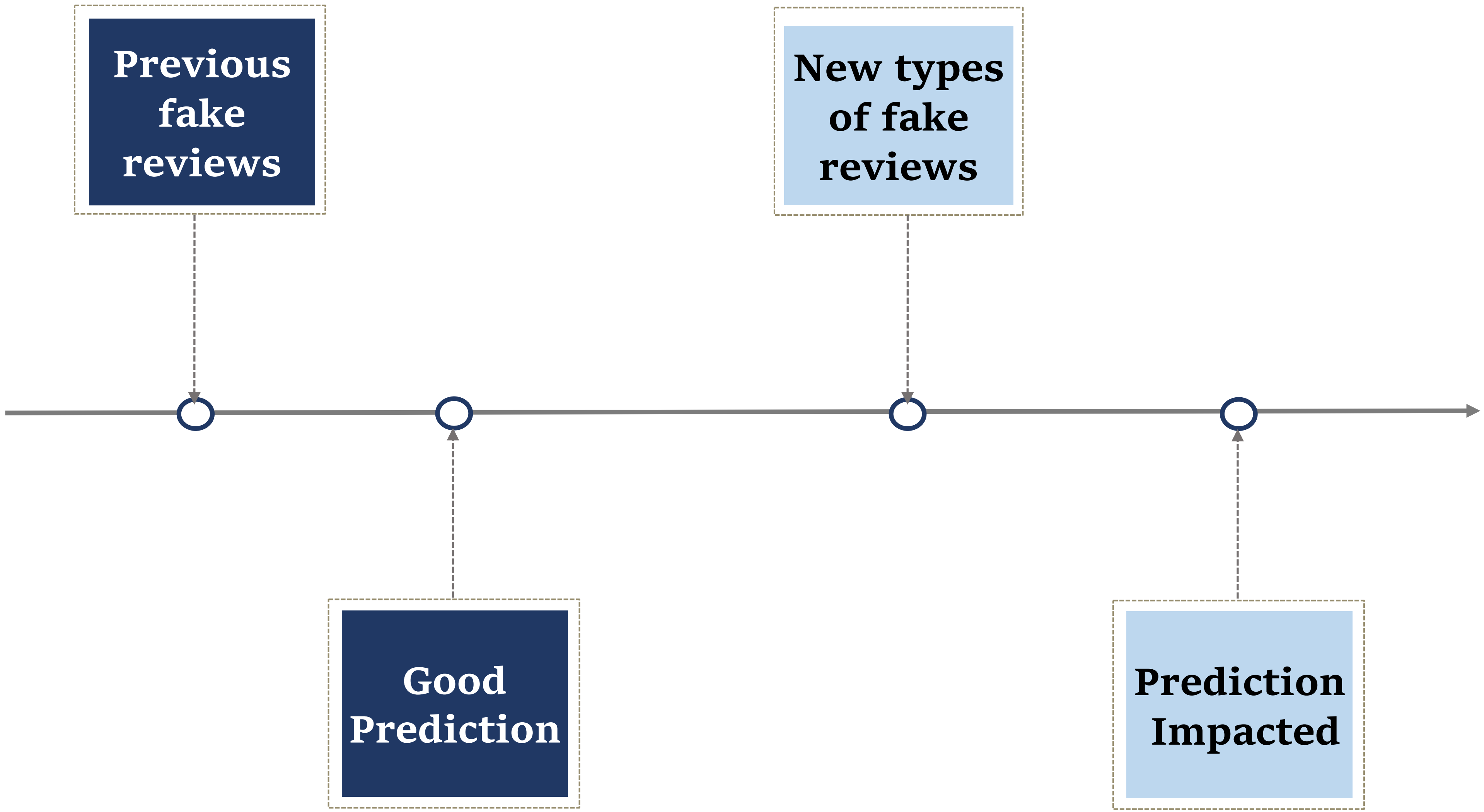
After Reweighting

Private Score	Public Score
0.40198	0.38391

Limitations: Training Data Quality



Limitations: Changes in Review Data Over Times



Improvements

- **Incorporate more additional features:**
- Metadata about the reviewer and product
- e.g., Age and gender about the reviewer
- e.g., Price and category of the product

- **Monitor and adjust the model's parameters**

- **Explore more ML techniques**

Business Analysis

Results

Outcomes

- Identified significant data imbalance
- Found critical features: *review_rating, number_of_photos, year, month, polarity, and subjectivity*

Performance

- Best model: DNN
- Best F1 Score: 0.4

Optimization Strategies

- Resampling
- Word Embeddings
- Reweighting
- Targeted Feature Engineering
- Parameter Fine-Tuning

Business Goals

Improve fake review detection

Boost model's effectiveness to identify and eliminate fake reviews

Authentic review section

Enhance the review section's authenticity

Assess impact of fake reviews

Discover impact of fake reviews on the platform's overall performance and profitability

Reasoning Behind Recommendations

Impacts on Sales	❖ **Increase** rapidly with fake reviews, but then **fall back** to original level	❖ **Decrease** with high percentage of fake reviews
Impacts on Ranking	❖ Fake reviews artificially **inflate** product rankings	❖ If depend more on fake reviews than quality, **product costs** will raise
Other Impacts	❖ Fake reviews lack factual info »» **unhelpful**	❖ Paying for fake positive reviews damages **product quality** and **consumer loyalty**

Implementations

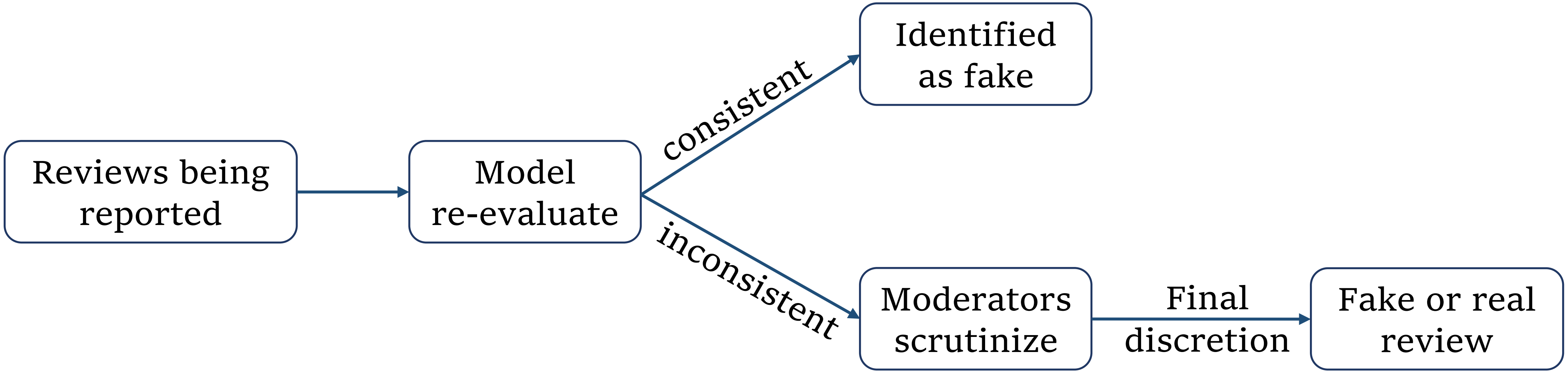
Implementations:

Fake review detected:

Remove

Fake review not detected:

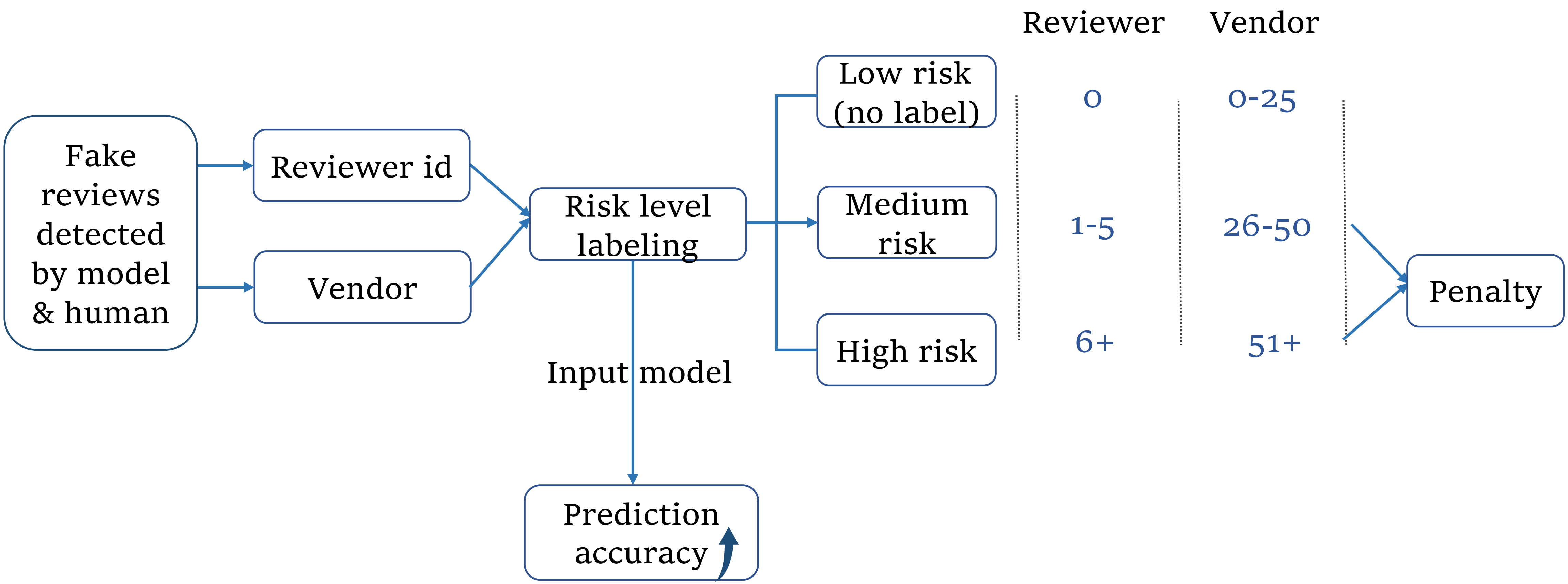
Introduce report system



Fake review report system

Implementations:

- Penalty mechanism: **risk labeling & penalty**



Conclusion

Presenter | Team SimonMiracle
Zhiwei Liang | zliang19@simon.rochester.edu
Yuxiao Liao | yliao21@simon.rochester.edu
Yitong Li | yli295@simon.rochester.edu
Xinyu Su | xsu9_Simon@simon.rochester.edu
Tingting Liu | tliu44@simon.rochester.edu