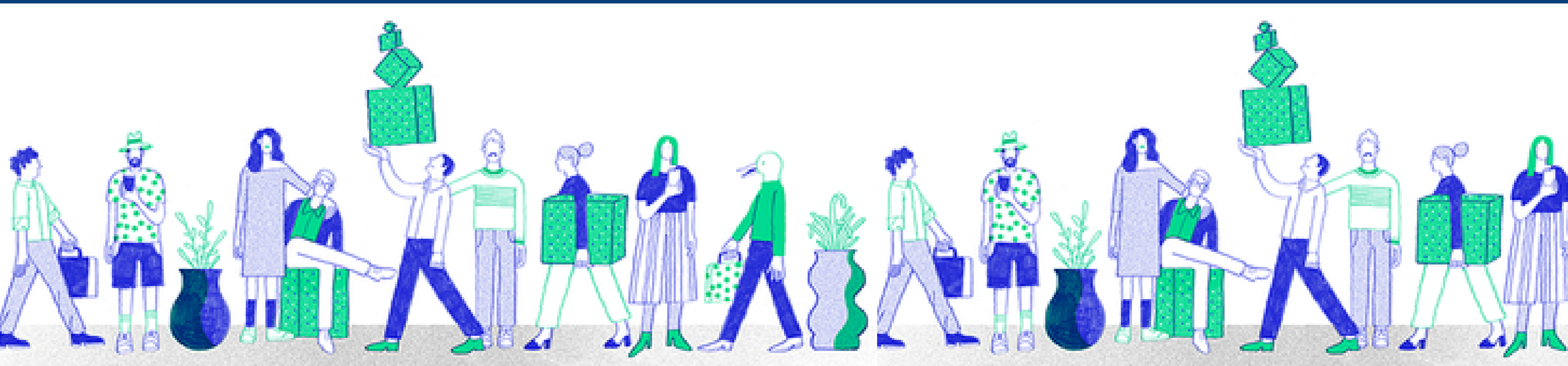
OLIST

Identifying potential repeat buyers

THE TEAM | Stakeholders Meeting



WHY REPEAT BUYERS MATTER

Stable revenue	Higher lifetime value (LTV)
Long-term growth	Loyal buyers refer others and boost organic growth
Better ROI for Olist	Repeat buyers already trust Olist, so they need less advertising to buy again

EDA

Datasets with duplicates: geolocation

Datasets with null values: reviews, products, orders

Key changes:

	Issues	cleaning
Geolocation	• Lat/Lng outside Brazil • noisy zip-level spread	• Removed invalid coordinates • Dropped zip prefixes with large spatial variance
Orders	Orders skewed right	• Filtered to 2016–2018 • Imputed missing delivery dates

EDA

Datasets with duplicates: geolocation

Datasets with null values: reviews, products, orders

Key changes:

	Issues	cleaning
Order_items	Price and freight outliers	• Filtered to 2016 - 2018 • Capped outliers at 99.9 percentile
payments	different payment preferences	• Changed not_defined /null to others payment type • Capped payment_installments at 99 percentile • Merged multi-method payments per order_id to simplify and preserve totals

FEATURE ENGINEERING OVERVIEW

Four feature buckets:

Customer Convenience

- Delivery cost
- Discounts
- Total spend
- Distance to seller
- Type of payment

Perceived value

- Free shipping
- Good deals

Product Appeal

- Photos
- Descriptions
- Product size/weight

Customer sentiment

- Review scores

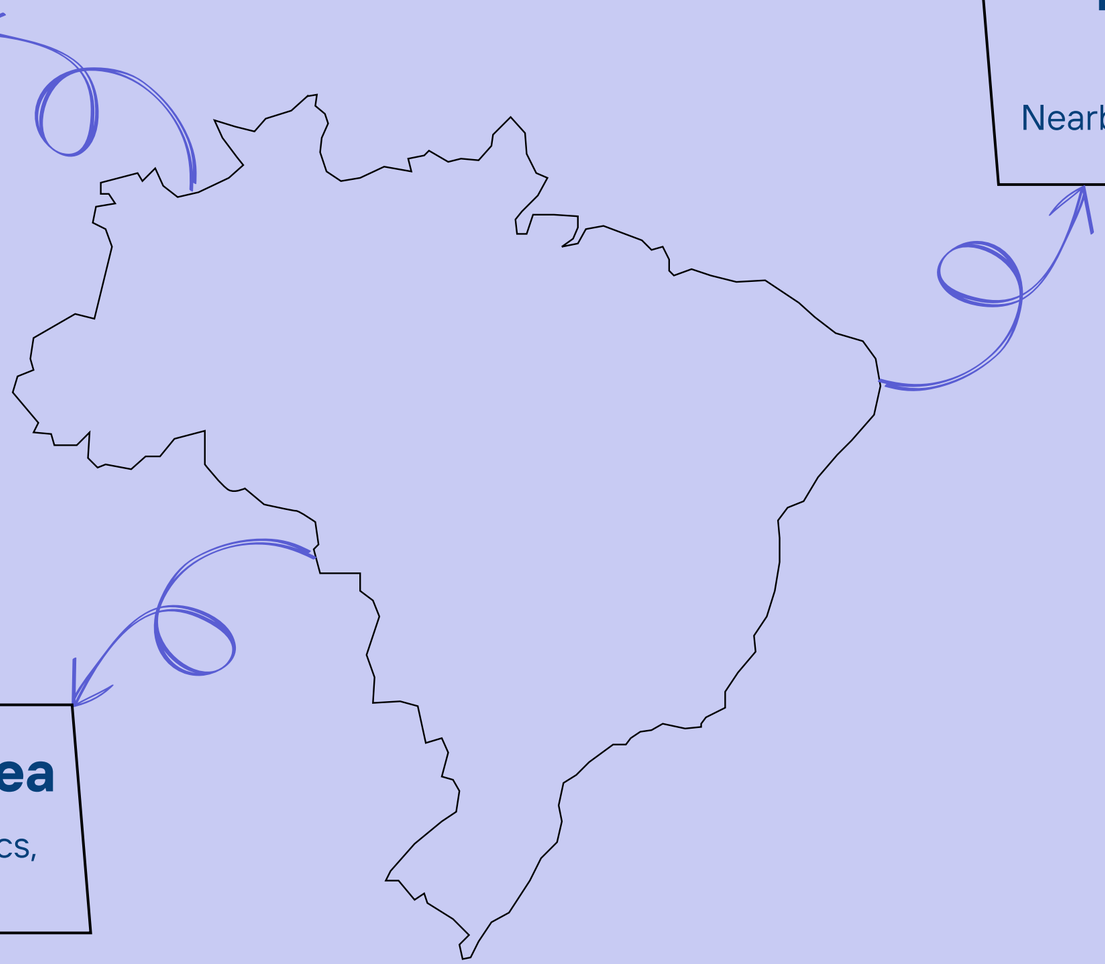
GEOGRAPHICAL PROXIMITY

Seller Performance

Trusted sellers = More repeat buyers

Buyer-Seller Distance

Nearby sellers = stronger buyer loyalty



High Customer Area

Urban areas = better logistics, more repeat buyers

DELIVERY & PRICE FEATURES

Delivery Speed vs. Expectation

Faster than expected = trust boost

Payment to Delivery Lag

Shorter wait = happier customer

Shipping Cost

Lower cost leads to higher satisfaction

Discount Usage

Promotions attract deal-seekers

Photos

Represents the number of images
More images → Higher buyer trust & confidence.

Product descriptions

Longer descriptions often reflect product clarity.
Reduces uncertainty → Increased satisfaction & reduced return rates.

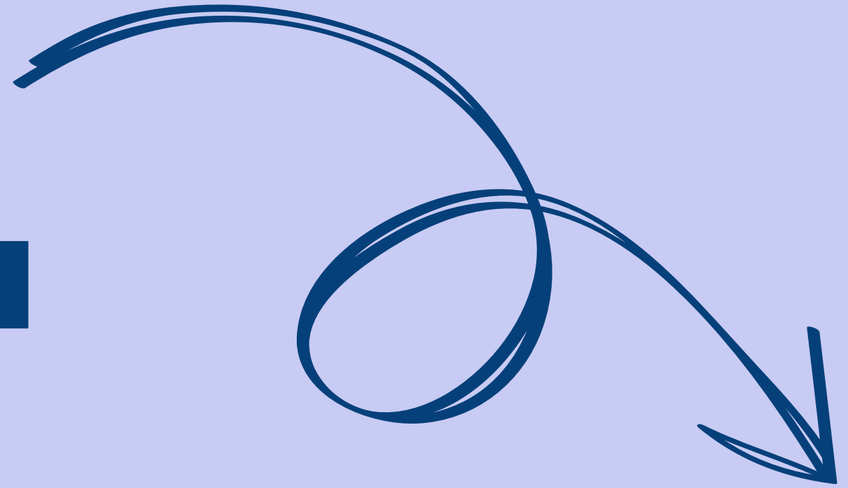
Product size/weight

Important for predicting freight value and order fulfillment success.
Larger or heavier items → higher shipping costs.

ORDER & PRODUCT CONTEXT

REVIEW SENTIMENT & VERIFICATION

Behind the Scenes



Sentiment analysis of comments

comparison with review score

review score verification

FEATURE DASHBOARD

Focus

Drivers of repeat-buyer behaviour.

Data Shown

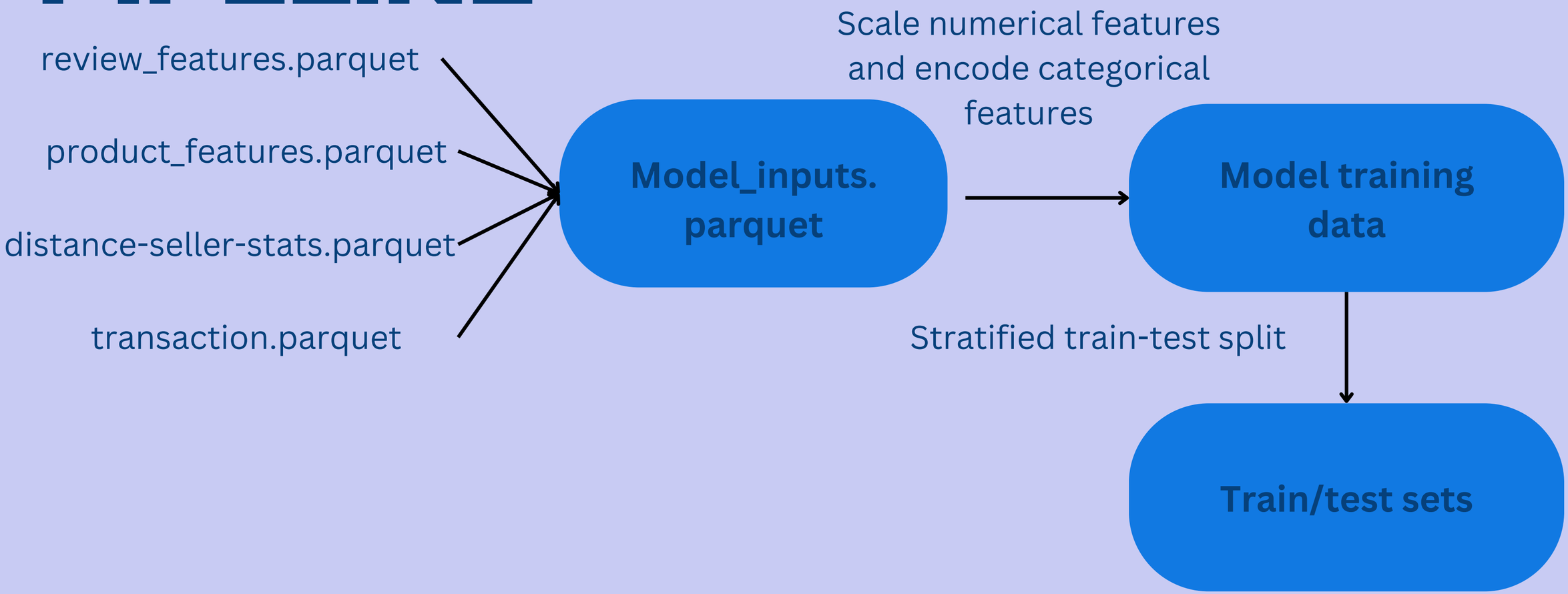
Key metrics like reviews, repeat rates, spend, and delivery.

Why is it useful?

Highlights where to improve listings and operations for loyalty.

MACHINE LEARNING PIPELINE

Ingestion



MACHINE LEARNING PIPELINE

Model selection *Accuracy

Logistic regression

69.2%

XGBOOST

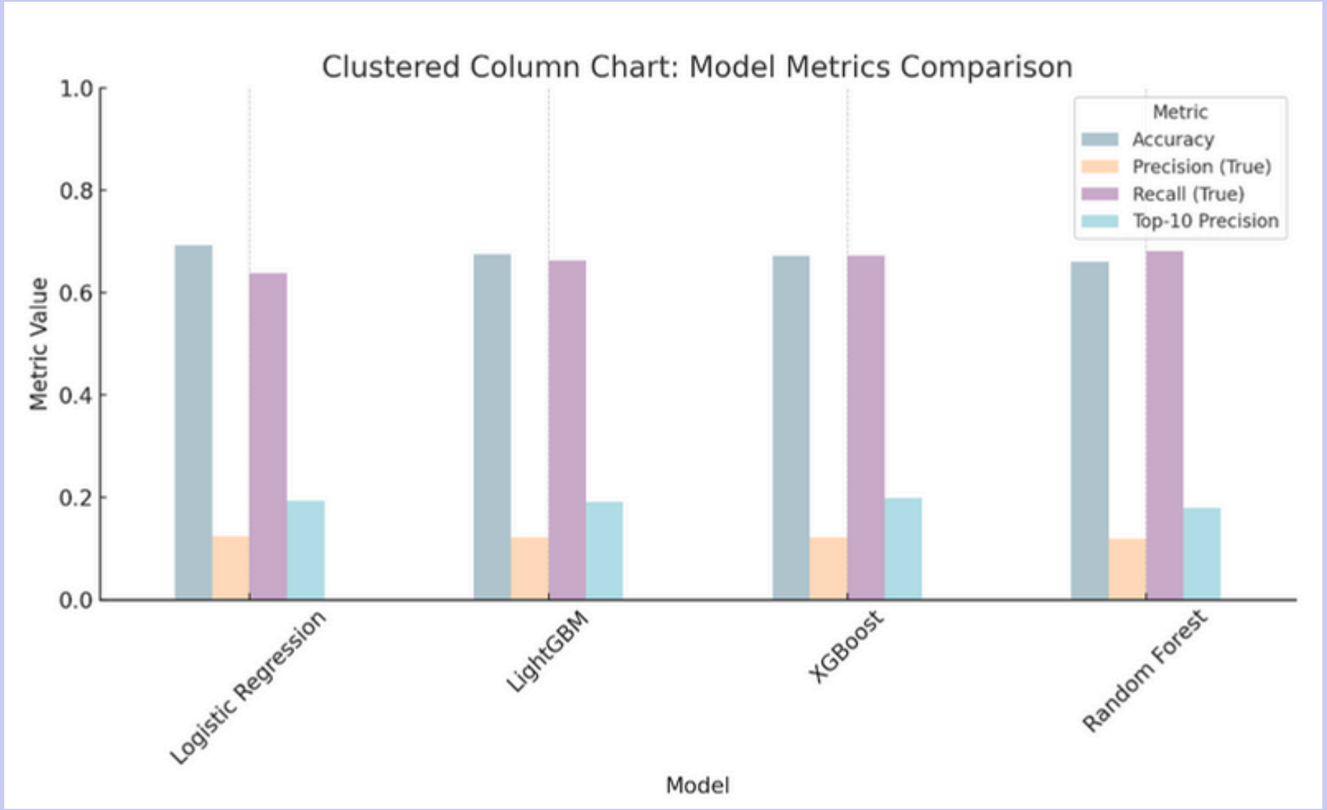
67.1%

LightGBM

67.5%

Random forest

66%



MACHINE LEARNING PIPELINE

Model selection

*Weighted avg Precision

Logistic regression

91.3%

XGBOOST

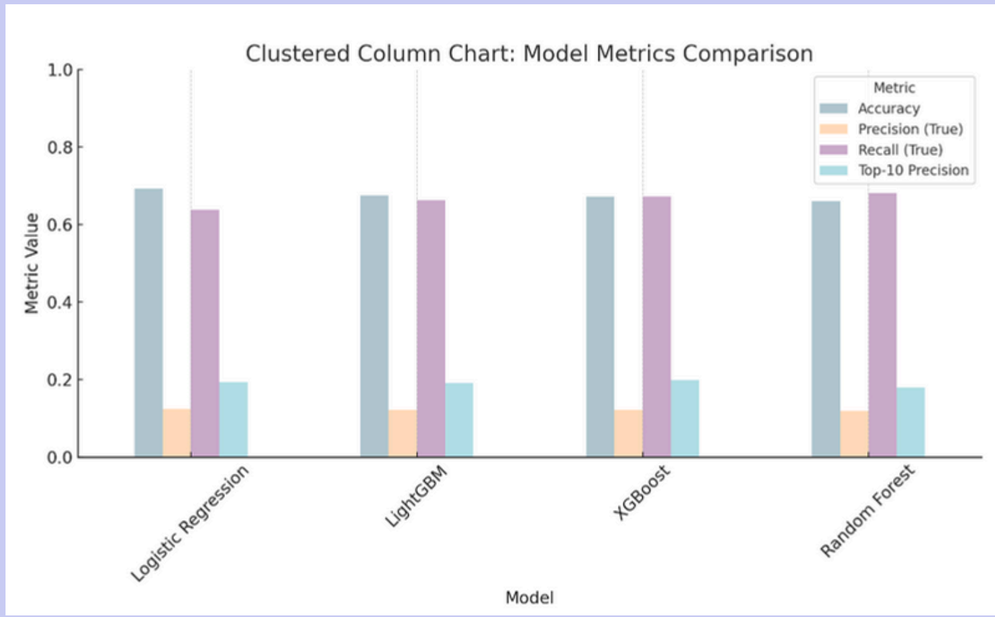
91.5%

LightGBM

91.4%

Random forest

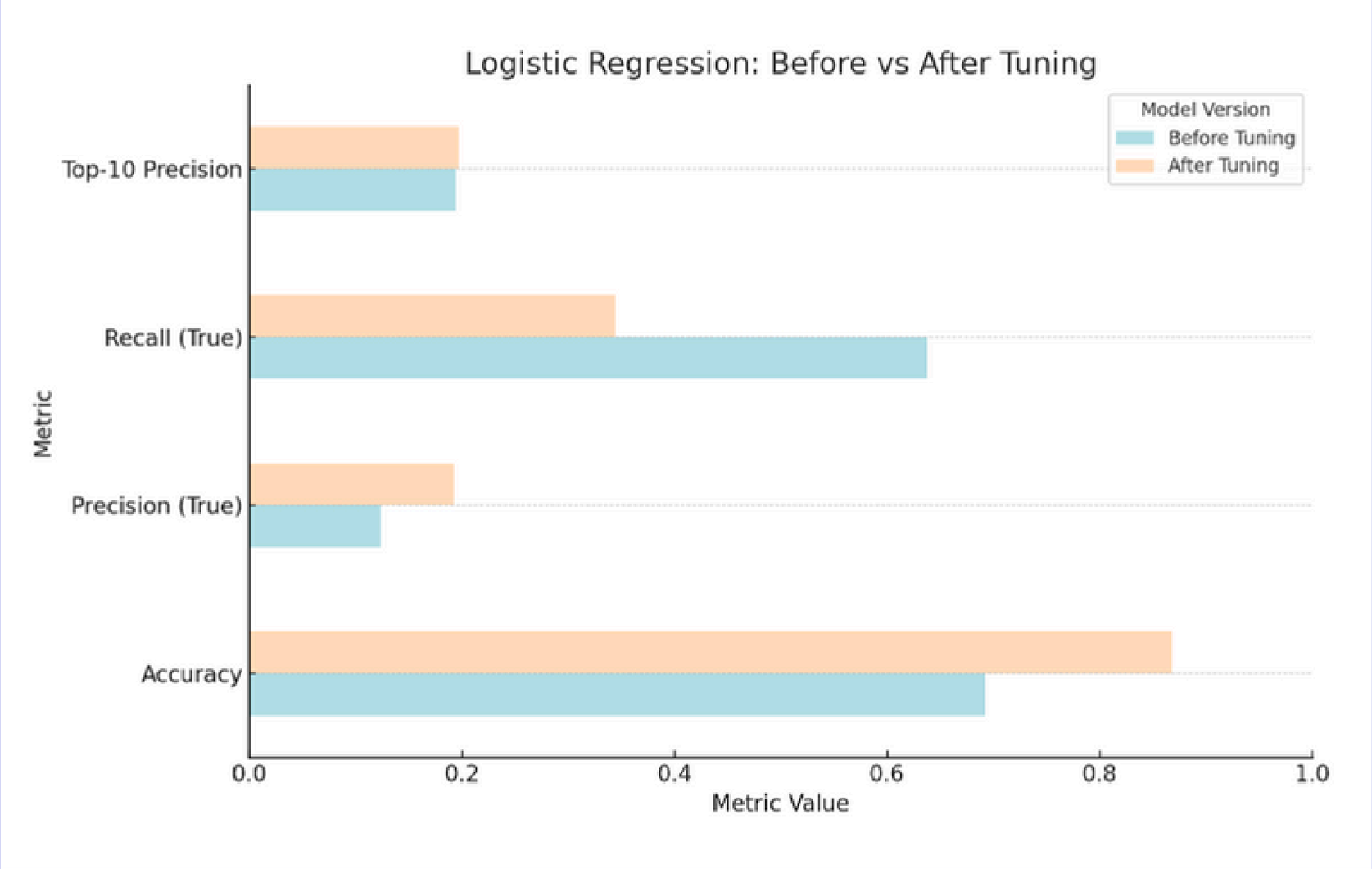
91.5%



KEY MODEL INSIGHTS

Fine-tuned Logistic Regression

Metrics used

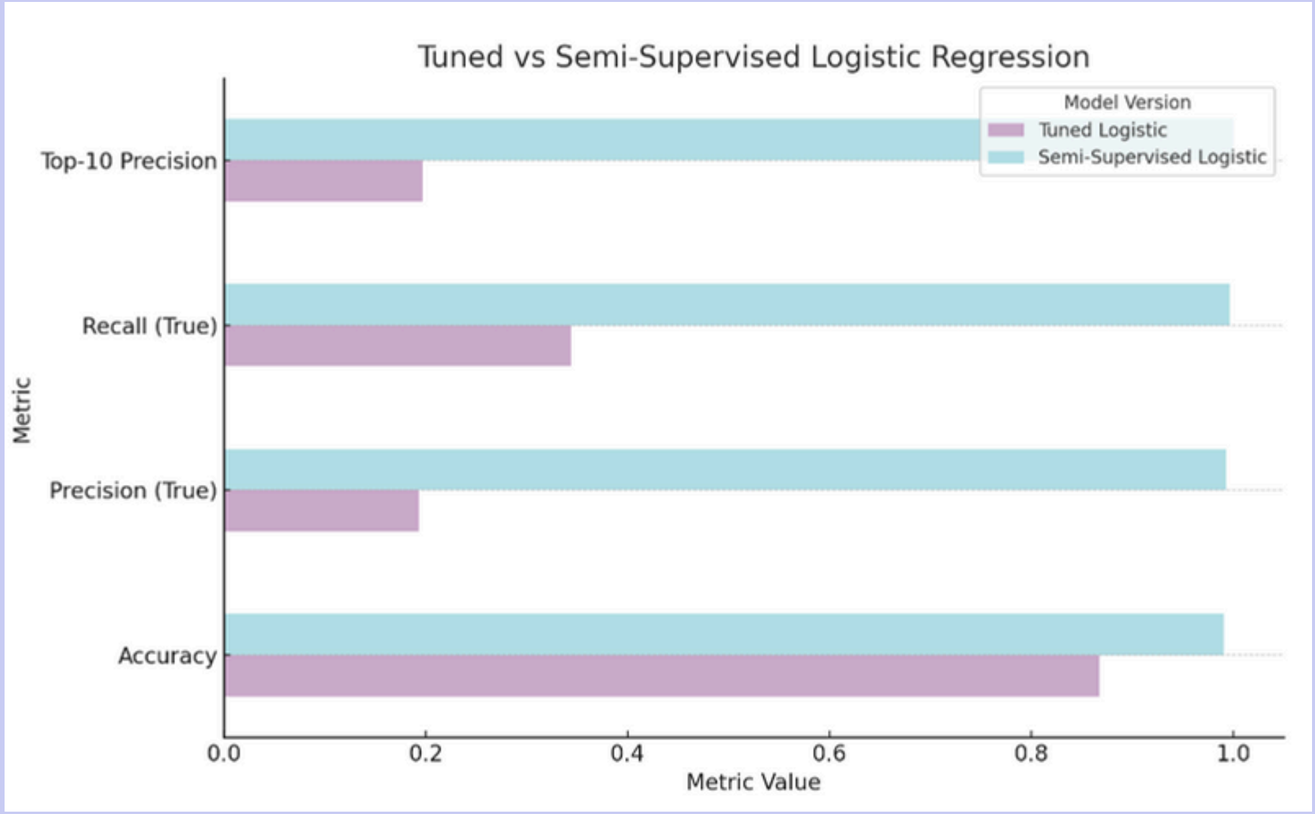


KEY MODEL INSIGHTS

Metrics used

Semi-supervised Fine-tuned Logistic Regression

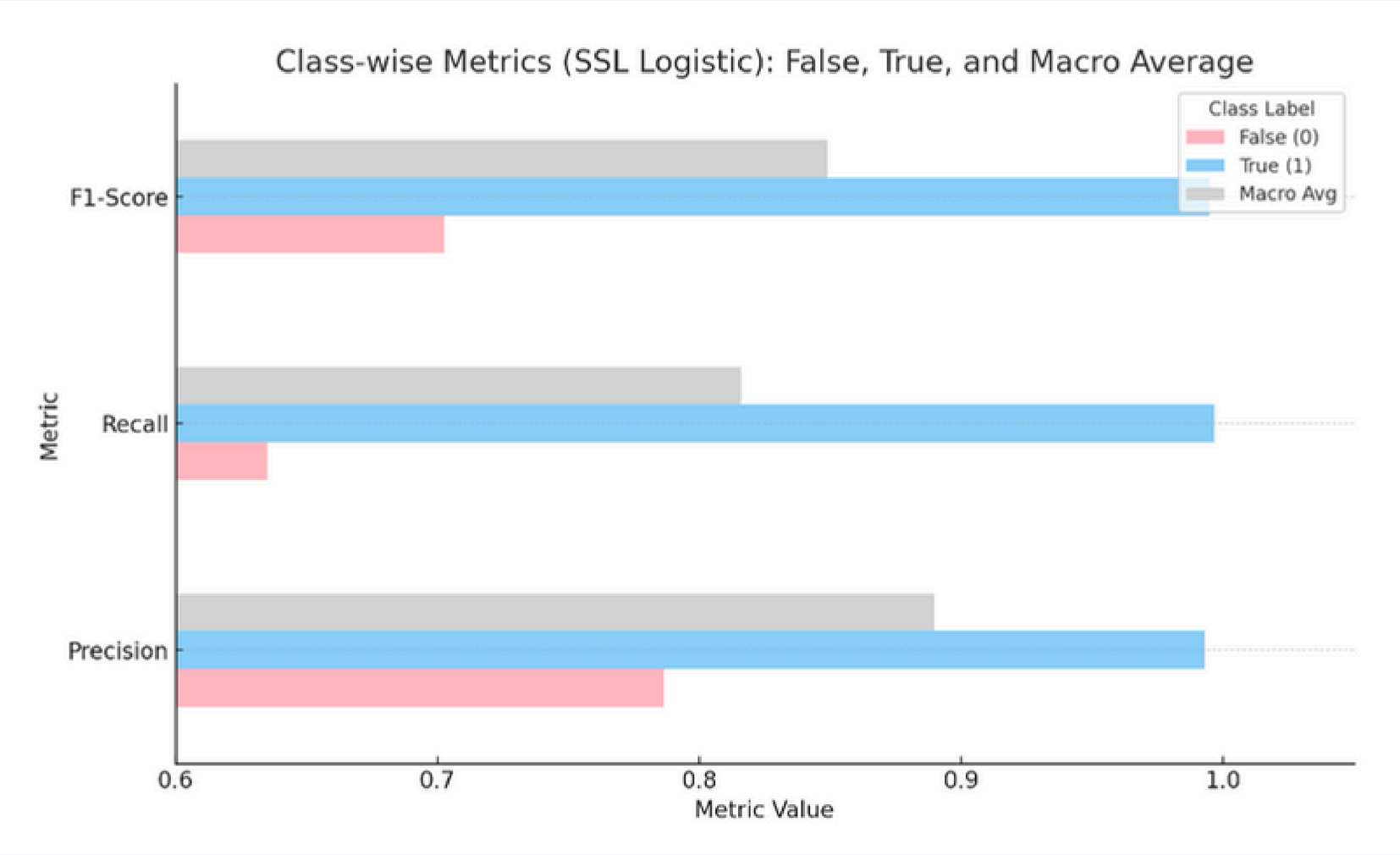
```
mask = (
    (df["review_score"] > 3)
    | (df["deli_duration_exp"] <= -7)
    | (df["voucher"] >= 0.3)
    | (df["total_spent"] >= df["total_spent"].quantile(0.8))
    | (df["product_category_name"].isin(top_categories))
)
return pd.DataFrame({"weak_positive": mask})
```

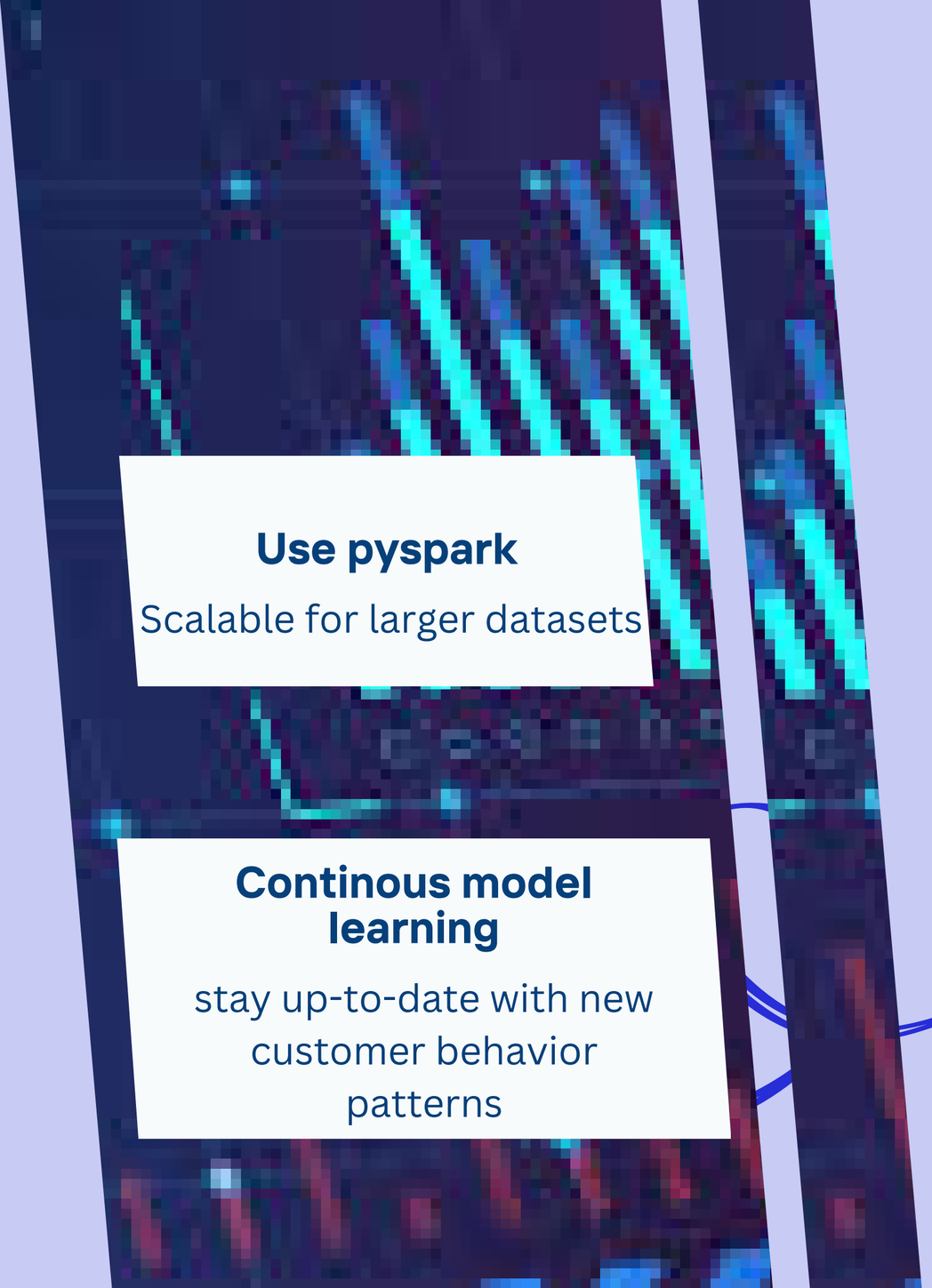


KEY MODEL INSIGHTS

Semi-supervised
Fine-tuned Logistic Regression

Metrics used





FURTHER ENHANCEMENTS

Use pyspark

Scalable for larger datasets

Continous model learning

stay up-to-date with new customer behavior patterns

THANK YOU & Q&A