MACHINE LEARNING

CUSTOMER CHURN ANALYSIS

PRESENTED BY:

Raoyi LI
Grace AIYEDOGBON
Yi-Chieh LEE
Jenny TING
Keer HUANG

BUSINESS PROBLEM & OBJECTIVE

Problems:

- Customer attrition erodes revenue and inflates marketing spend.
- Acquiring a new buyer can cost 5–25 times more than retaining an existing one

Objective:

- Classify customers into churners vs. non-churners
- Use behavioral and transactional data to guide retention strategies.

CRISP-DM PHASE 1: BUSINESS UNDERSTANDING

- Maximize net lifetime value by reducing voluntary churn.
- A churn event removes future cashflows and forces fresh acquisition.
- Management sets an intervention threshold:

-If predicted risk exceeds threshold → trigger CRM retention action.

-Threshold chosen via cost–benefit analysis for profit optimization.

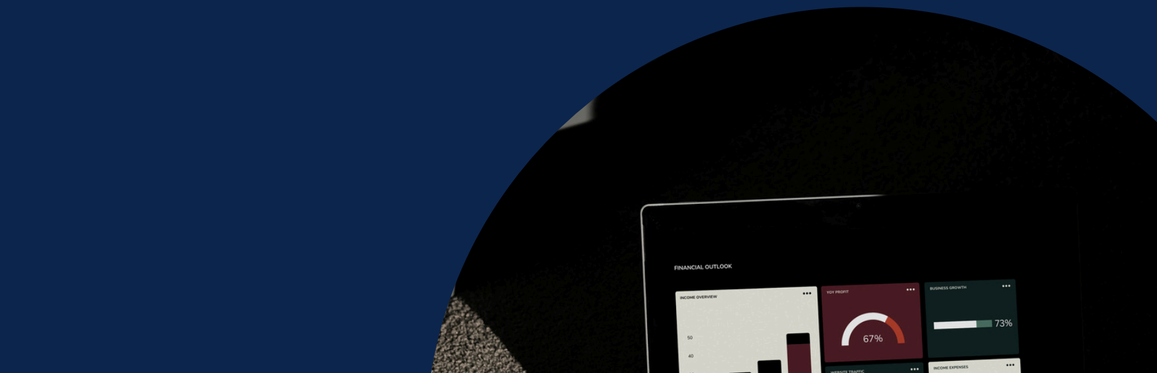
CRISP-DM PHASE 2: DATA UNDERSTANDING

- Dataset: 5630 observations, 17% churn
- Mix of static (age, region) and transactional (tenure, order count) features
- Class imbalance motivates metrics like F1-score and AUC-ROC
- Continuous skew suggests log/quantile transforms for modeling

DATA PREPROCESSING - KEY STEPS

1. Initial Data Inspection
2. Missing Value and Duplicate Handling
3. Data Type Conversion
4. Categorical Feature Encoding

5. Numerical Feature Scaling
6. Standardization (StandardScaler)
7. Normalization (MinMaxScaler)
8. Cleaned Data Export



LOGISTIC REGRESSION

PURPOSE OF THE MODEL

Logistic Regression is used to predict whether a customer will churn or not, based on behavior and transaction history.

WHY THIS MODEL?

It's fast, transparent, and shows if a customer characteristic increases or decreases churn risk.

WHAT CAN WE KNOW FROM THE MODEL?

- It provides a clear yes/no output for each customer: churn (1) or stay (0)
- Shows how customer attributes (e.g. inactivity, low satisfaction) influence churn risk
- Helps management see which behaviors are early warning signs

LOGISTIC REGRESSION

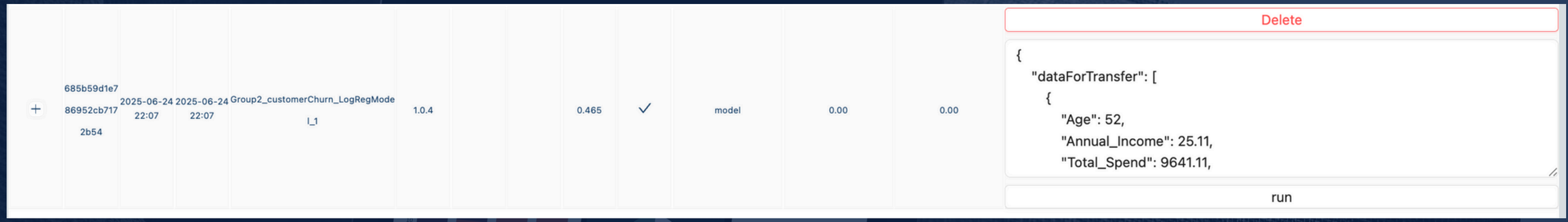
KEY RESULTS

Metric	Result	Meaning
Accuracy	46.50%	The model correctly predicted churn or no-churn for 46.5% of all customers, yet it is not good enough to use by itself.
Recall	66.00%	The model caught 66% of the customers who actually left —> useful for finding people at risk.
Precision	49.60%	Out of all the customers the model predicted would churn, 49.6% really did.
F1 Score	0.567	Average prediction error, which shows a balance between finding leavers and avoiding mistakes. 0.567 means the model does a decent job finding churners but also makes some mistakes. —> Useful for early warnings, but not reliable enough for automatic decisions.

LOGISTIC REGRESSION

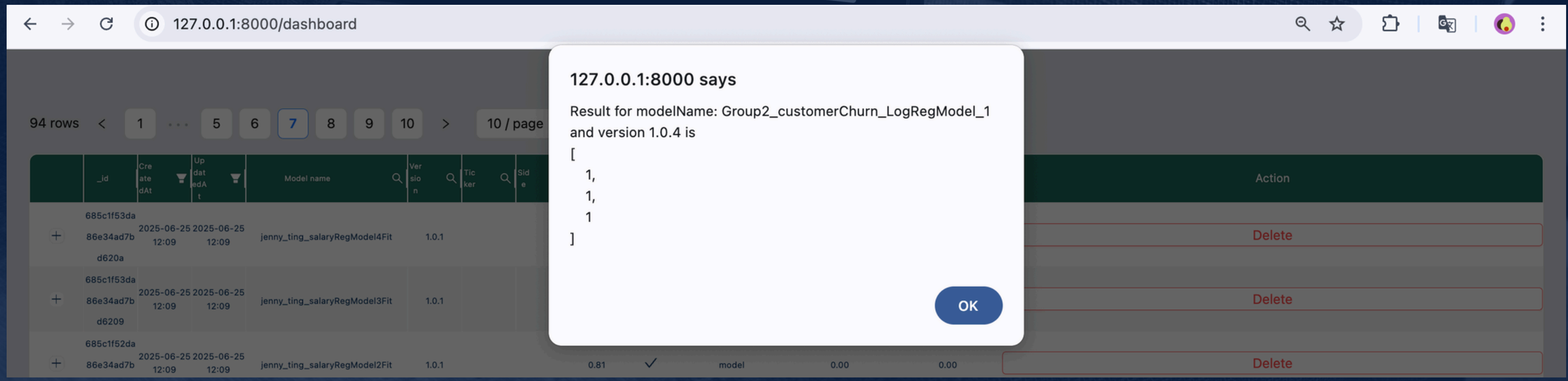
MODEL DEPLOYED & TESTED ON JRJMODEL DASHBOARD

The Logistic Regression model was successfully registered to the JRJModel dashboard and can now be tested using live input data.



Logistic Regression model predicts the customer will churn.

-> The model is now ready to be used as a real-time service to predict churn for future customers.



DECISION TREE

PURPOSE OF THE MODEL

A Decision Tree is a transparent, rule-based classifier that identifies churn risk through a series of simple “if-then” splits on customer attributes. Its interpretability lets us trace exactly why any customer is flagged, enabling clear, actionable retention rules.

FEATURE IMPORTANCE

- Age
- Total_Spend (last year’s total purchase)
- Years_as_Customer (tenure)
- Num_of_Purchases (purchase frequency)
- Average_Transaction_Amount
- Last_Purchase_Days_Ago (recency)

DECISION TREE -MODEL 1 (INITIAL)

KEY RESULTS

Metric	Result	Meaning
Accuracy	56.00%	Overall fraction of correct predictions. The un-tuned tree labels churn vs. non-churn correctly just over half the time.
Recall	58.00%	Of all customers who actually churned, 59 % were flagged. The model is catching roughly 6 out of 10 true churners.
Precision	58.00%	Of those predicted to churn, 58 % truly did. **About 4 in 10 "at-risk" alerts are false positives** at this stage.
F1 Score	59.00%	The harmonic mean of precision and recall. A 59 % F1 reflects a modest balance between detecting churners and limiting false alarms.

DECISION TREE - MODEL 2 (TUNED)

KEY RESULTS

Metric	Result	Meaning
Accuracy	49.50% ↓ -6.5%	Overall proportion of correct predictions—this model correctly classifies about half of all customers as churners or non-churners.
Recall	**77.10%** ↑ +19%	Of all customers who actually churned, 77.1 % were correctly identified. High recall means we catch most at-risk customers.
Precision	51.30% ↓ -6.7%	Of those predicted to churn, 51.3 % truly did. Roughly half of our “at-risk” alerts are accurate, **the rest are false alarms which means potential waste of resources**
F1 Score	**61.60%** ↑ +2%	The harmonic mean of precision and recall, balancing both detection of churners and minimization of false positives.

RANDOM FOREST

PURPOSE OF THE MODEL

Random Forest is a robust, ensemble-based method that handles both numerical and categorical features well, avoids overfitting through bagging, and is highly interpretable through feature importance.

FEATURE IMPORTANCE

- Annual Income
- Total Spend
- Average Transaction Amount
- Last Purchase Days Ago
- Num of Purchases
- Age
- Years as Customer
- Num of Returns

- Segment customers into income tiers and test tailored pricing, exclusive offers, or loyalty perks for high-income or low-income segments.
- Prioritize retention offers to high spenders (VIP programs) or upsell low spenders to increase lifetime value.

RANDOM FOREST - MODEL 1 (INITIAL)

KEY RESULTS

Metric	Result	Meaning
Accuracy	54.00%	The proportion of all correct predictions. The model correctly predicted churn 54% of the time.
Recall	63.20%	The ability to identify actual churners. Of all customers who actually churned, 66% were successfully identified by the model.
Precision	55.80%	The proportion of churn predictions that were correct. Among all customers predicted to churn, 54.7% actually did.
F1 Score	0.592	The harmonic mean of precision and recall. This score balances both metrics, and a value of 0.598 indicates the model achieves a fair compromise between capturing churners and minimizing false alarms.

RANDOM FOREST - MODEL 2 (TUNED)

KEY RESULTS

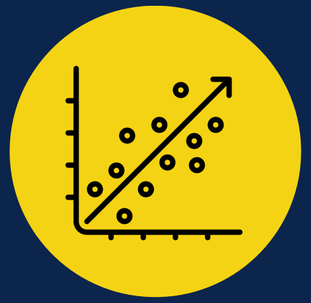
Metric	Result
Accuracy	53.00% ↓
Recall	**72.20%** ↑
Precision	54.00% ↓
F1 Score	**62.00%** ↑

BASELINE MODEL COMPARISON

KEY RESULTS

Metric	Accuracy	Recall	Precision	F1 Score
Logistic Model	46.50%	66.00%	49.60%	56.70%
Decision Tree	56.00%	59.00%	58.00%	58.24%
Random Forest	54.00%	63.20%	55.80%	59.30%

INSIGHTS & ETHICAL CONSIDERATIONS



LOGISTIC REGRESSION

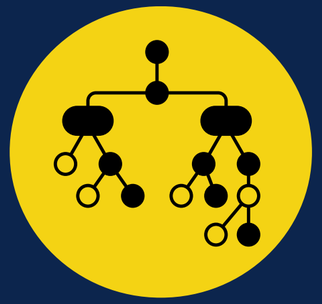
Key Insight: Identified 66% of churners, showing it is effective for early churn detection. However, only 49.6% precision may trigger false alarms -> Need to combine it with stronger models.

Ethical Consideration: The linear nature makes it susceptible to amplifying existing biases in training data, potentially leading to unfair or ineffective targeting.

RANDOM FOREST

Key Insight: Its precision improved to 55.8%, indicating fewer false alarms compared to the baseline.

Ethical Consideration: The model can perpetuate data biases if not carefully managed, potentially leading to unfair outcomes or an over-reliance on automated predictions.



DECISION TREE

Key Insight: Hyperparameter tuning and feature selection did not improve accuracy performance, however, it improves for recall by 20% meaning we capture at least 77% of churner.

Ethical Consideration: It carries the ethical risk of perpetuating data biases.

TOP MODEL - IMPROVE CHURN RATE

DECISION TREE -MODEL 2 TUNED

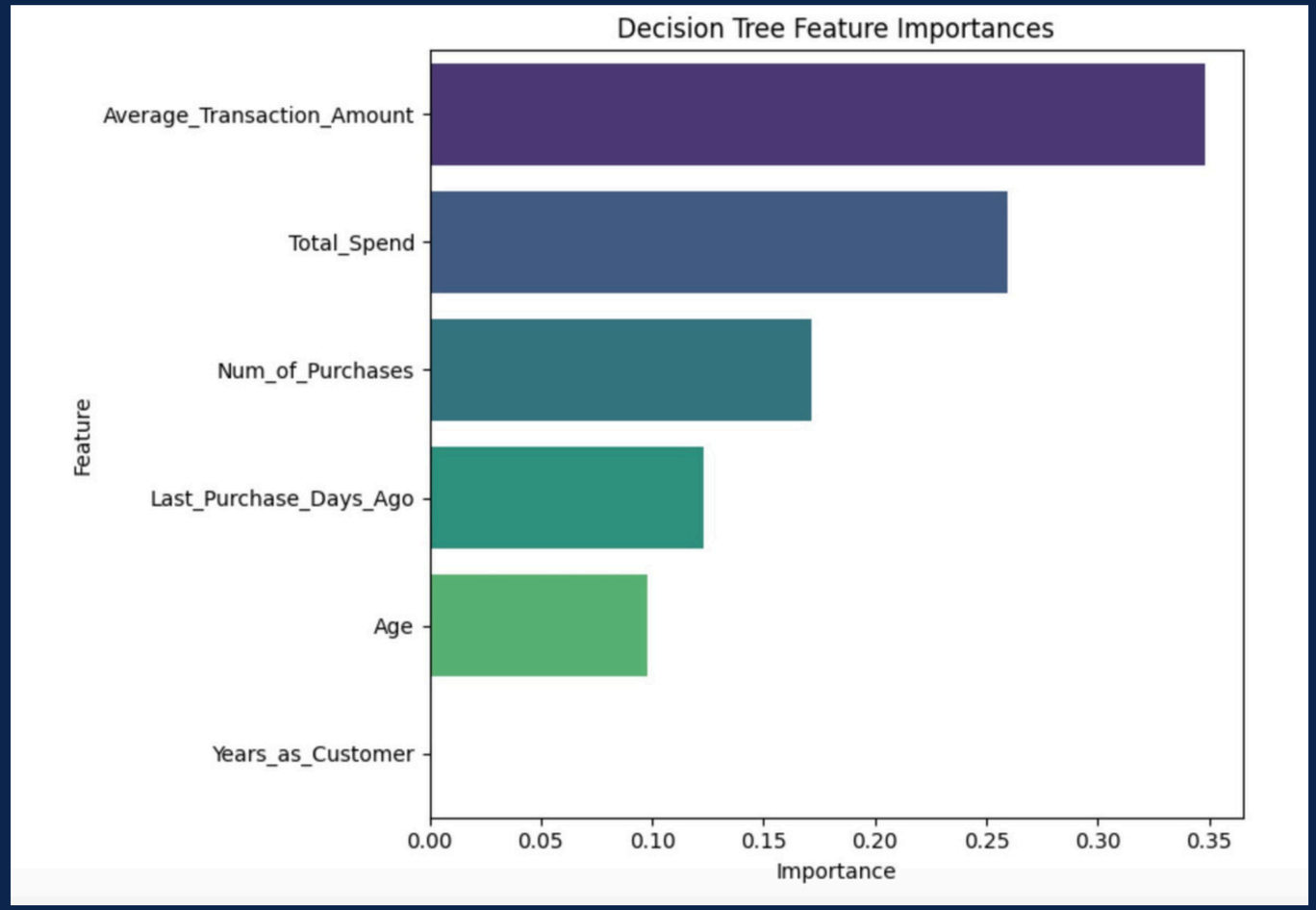
This model achieved the highest Recall at 77%, outperforming both the Baseline of Logistic Regression and the tuned Random Forest(72%) The model’s inherent interpretability is a significant advantage of the purpose for proactive retention efforts.

Consider low-cost outreach (email, SMS) broadly, reserving high-cost tactics (calls, offers) for top-risk segments. Adjust the decision threshold to balance recall and precision based on retention budget.

TOP MODEL - IMPROVE CHURN RATE

DECISION TREE - ADDITIOAL INSIGHT

PRECISION is at 51%, about half of predicted churners actually leave, indicating significant outreach to loyal customers, which may raise marketing costs.



WE WANT TO SAY

THANK YOU

FOR YOUR ATTENTION