

Propensity Modeling for Subscription Adoption

1. Executive Summary & Key Results

Business Question: "Which existing Stripe merchants are the *highest-value candidates* for cross-selling the Subscriptions product, based on their behavioral similarity to our current successful users?"

Key Findings

- **Adoption Drivers:** The strongest predictors of subscription adoption are Merchant Industry (specifically Software/Digital Goods) and Platform Tenure.
- **Greenfield Opportunity:** Correlation analysis reveals that usage of "Checkout" and "Subscriptions" is uncorrelated, indicating that cross-selling will generate net-new revenue rather than cannibalizing existing transaction volume.
- **Model Performance:** The prediction model achieved an AUC of 0.79, successfully distinguishing high-potential leads from the general population.

Recommendations

- **Target List:** Deploy sales resources immediately to the 1,000 high-propensity merchants identified in the attached target_merchants.csv.
- **Segmented Outreach:** Tailor marketing messaging for the Software and Digital Goods sectors, as they show a naturally higher product-market fit (14.2% feature importance).

Expected Impact

- **Efficiency:** By focusing only on the top 1,000 leads (vs. random outreach), we expect a significantly higher conversion rate and lower customer acquisition cost.
 - **Revenue:** Targeting "Whales" (high TPV merchants) ensures that each conversion brings substantial incremental processing volume.
-

2. Data Overview & Stated Assumptions

Data Sources

- **payments Dataset:** Daily transaction volume (in cents) for Subscriptions, Checkout, and Payment Links (2041–2042).
- **merchants Dataset:** Firmographic data including industry, country, business size, and signup date.

Assumptions

- **Data Representativeness:** We assume the provided data is a random sample of future merchants.
- **Dataset Selection:** We assume the most recent upload of the merchants dataset (merchants.csv) is the updated/up-to-date data.
- **Merchant ID Corruption:** We assume that some Merchant IDs are corrupted due to lossy compression into scientific notations
- **Target Definition:** We define a "Subscriber" as any merchant with subscription_volume > 0.
- **Currency:** All transaction volumes were converted from cents to dollars for analysis.

Metric Proposal: Success Metrics

- **North Star Metric:**
 - a. **Incremental TPV:** The total volume processed through Subscriptions by the new cohort. The primary goal is to increase the total volume processed through Stripe, not just the number of users; we want to validate that we are not cannibalizing existing volume
 - b. **Conversion Rate (CVR):** The percentage of the targeted 1000 merchants who start using the Subscription product within 30 days. We want to measure our immediate campaign effectiveness
 - c. **Retention Rate:** Evaluate churn rates between multi-product users (comparison between multi-product users vs. single-product users). We can test if multi-product adoption increases platform stickiness.
 - **Guardrail Metric:** Checkout Retention & Payment Link Retention. We must monitor that aggressively pushing Subscriptions does not cause merchants to stop using other products (churn).
-

3. Data Preparation

Data Quality & Cleaning (etl.py, data_prep.py)

Data Cleaning Summary

Metric	Count / Value	Notes
Original Merchants	23,627	Source metadata

Corrupted IDs Removed	130	Excluded merchants with corrupted IDs
Valid Matches	23,497	Merchants with valid payment history
Negative Tenure Dropped	1,099	Removed rows with invalid start dates
Final Modeling Set	22,392	High-quality data used for training

- **ETL/Checkpoint (.parquet):** `pd.read_excel()` is slow with .xlsx files because python has to unzip the file first then parse the xml cell by cell; so we do a quick ETL step to read the .xlsx file once then save it to a high-speed parquet file
- **The "Ghost ID" Corruption:** Identified a critical data quality issue where ~0.5% of Merchant IDs were corrupted into scientific notation (e.g., 2.72E+98) in the source files. Using a strict regex filter, these 130+ unrecoverable records were excluded to prevent volume attribution errors.
- **Temporal Logic:** Removed rows with negative tenure (where `first_charge_date` was missing or invalid) to ensure logical consistency.

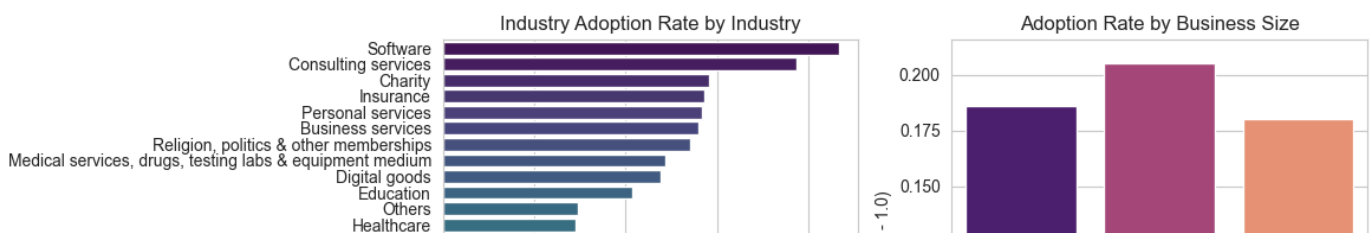
Feature Engineering (data_eng.py)

- **Aggregated Profile:** Grouped 1.5M daily transaction rows by merchant ID to construct "Merchant Profile" containing metrics like Total TPV, Volatility (Standard Deviation), and Tenure.
- **Leakage Prevention:** Explicitly excluded subscription-related metrics from the predictor set to prevent Target Leakage.

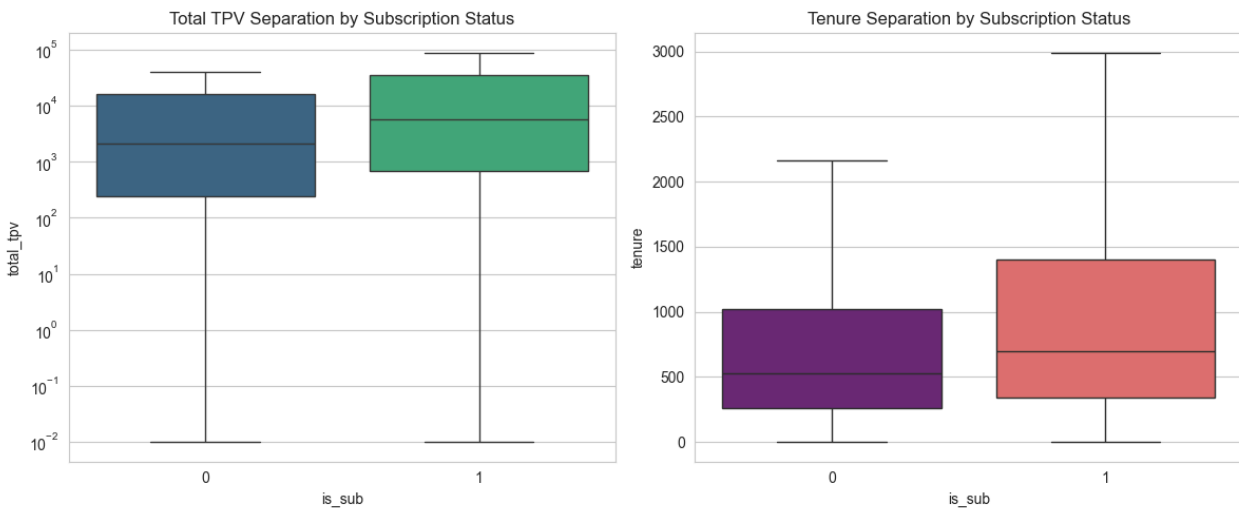
4. Approach and Rationale

Exploratory Data Analysis (EDA.py)

- **Outliers:** checked for outliers for our quantitative attributes & identified top merchants for data integrity. By confirming that `total_tpv` follows a Power Law (data made up with mostly small merchants and a few big ones), we know that the data is skewed
- **Segmentation Analysis:** examined adoption rate by Industry and also by business size achieve strategic focus



- **Separation Checks:** compared distributions of tenure and total_tpv for subscribers and non-subscribers to test attributes' validity to be used as a predictor.

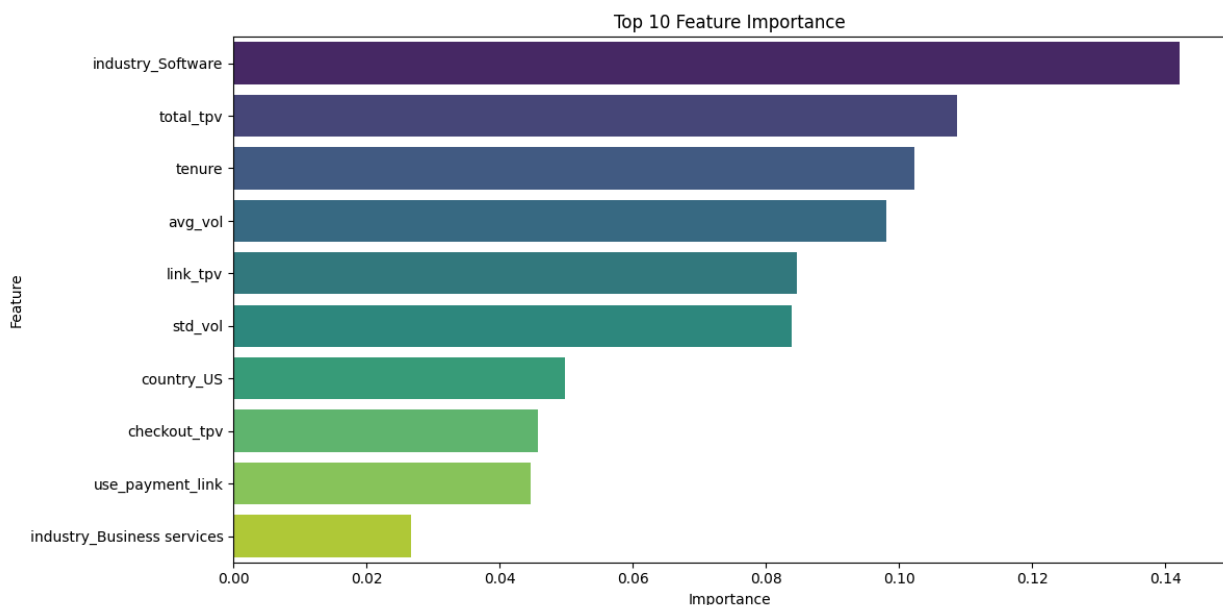


- **Correlation Matrix:** identified a near zero correlation between Checkout and Subscription usage, indicating that Checkout users do not convert themselves into Subscription users and that is where sales/marketing campaigns come in. It also provides a multicollinearity check before feeding inputs into our classifier.

Methodology (model.py)

Random Forest Classifier I selected a Random Forest model over linear alternatives for the following reasons:

1. **Handling "Power Law" Data:** Financial transaction volume is highly skewed. Random Forest handles these outliers ("Whales") naturally without requiring extensive log-transformations.
2. **Non-Linear Patterns:** EDA showed that Tenure has a non-linear relationship with adoption (a "maturity threshold"). Tree-based models capture these thresholds automatically.
3. **Explainability:** The model provides Feature Importance scores, allowing us to explain



why a merchant was selected to the sales team.

Validation Strategy

- **Metric:** Optimized for AUC (Area Under the Curve) rather than Accuracy. Since the dataset is imbalanced (~18% subscribers), AUC is the superior metric for ranking/prioritization tasks.
- **Statistical Testing:** Validated key drivers using the Mann-Whitney U Test (for Tenure) and Chi-Square Test (for Business Size and Industry) to ensure observed differences were statistically significant ($p < 0.05$).

Metric	Value
Tenure Effect (Mann-Whitney U)	48341073.0
p-value	< 0.001
Median Tenure (Subscribers)	695 days
Median Tenure (Non-Subscribers)	531 days
Merchants using Checkout and Payment Link	303
Correlation - Checkout and Payment Link	0.0152
Merchants using Checkout and Subscription	745
Correlation - Checkout and Subscription	0.0082
Merchants using Link and Subscription	1070
Correlation - Link and Subscription	0.0070

5. Results

Model Performance

- **AUC Score:** 0.79 (Strong predictive power).
- **Precision (Positive Class):** 0.69. When the model predicts a lead is high-value, it is

correct 69% of the time, minimizing wasted sales calls.

Key Drivers of Adoption

Rank	Feature	Importance	Interpretation
1	Industry: Software	14.2%	Software companies have the highest natural fit.
2	Total TPV	10.9%	Larger businesses are more likely to need recurring billing.
3	Tenure	10.2%	Merchants active >1.5 years are the "sweet spot."

Target List

- Generated a prioritized list of 1,000 merchants (see the separate attached file)

Top 5 Merchants from Target List

merchant	propensity
a0a57ca4	0.6350179062042544
8c48700b	0.5731708692543314
dd0e676d	0.5697216475382721
a4d6eea2	0.56783578008415
1795ecea	0.5668790313178951

Note: the top merchant has the propensity score ~0.64, since the baseline adoption rate is only ~18%, a score of 0.64 represents a **3.5x lift** over the average merchant

6. Next Steps & Future Considerations

Limitations & Trade-offs

- **Snapshot Bias:** The current model uses a static snapshot of "Total Volume." It does not account for recent acceleration (e.g., a merchant whose volume doubled last month).
- **Cold Start:** The model requires transaction history to work, so it cannot score brand-new signups (Tenure < 30 days).

Future Improvements

- **Time-Series Features:** With more resources, I would engineer "Trend" features (e.g., Month-over-Month Growth Rate) to catch merchants before they peak.
 - **A/B Testing:** To measure lift more effectively, I would recommend conducting an A/B test where the null hypothesis could be "There is no scientifically significant difference in subscription adoption rates between High-Propensity merchants who receive sales outreach and those who don't."
 - **Look-alike Expansion:** Incorporate "Product Description" NLP data to find software companies that might be misclassified as "Business Services."
-

7. Summary

- Cold outreach is inefficient, the team needs a compact list of "lookalike merchants" who closely resemble our existing successful subscribers.
 - Developed a Random Forest Classifier to handle the skewed transaction data to target 1000 high-propensity merchants based on signals like platform tenure, total volume and industry fit.
 - This approach demonstrated an AUC score of 0.79 and precision of 69% which indicates strong predictive power and high-confidence leads
-

8. What I Learned

- **Data Integrity is Paramount:** Finding the "Scientific Notation" corruption taught me that even "clean" datasets can have hidden traps. A model built on those corrupted IDs would have been worthless.
- **Statistical rigor supports ML:** Using the Mann-Whitney U test gave me confidence that the "Tenure" signal was real before I even trained the model.
- **Business Context:** I learned that "Prediction" is not enough; "Ranking" (via AUC) is what actually helps a sales team prioritize their day.
- **Biggest Surprise:** I expected the merchants with the highest volumes would be the sole dominant driver, however, it turned out that while total transaction volume matters, platform tenure was equally critical.