

Stripe Data Science Internship Project Report: Identifying Subscription Product Targets

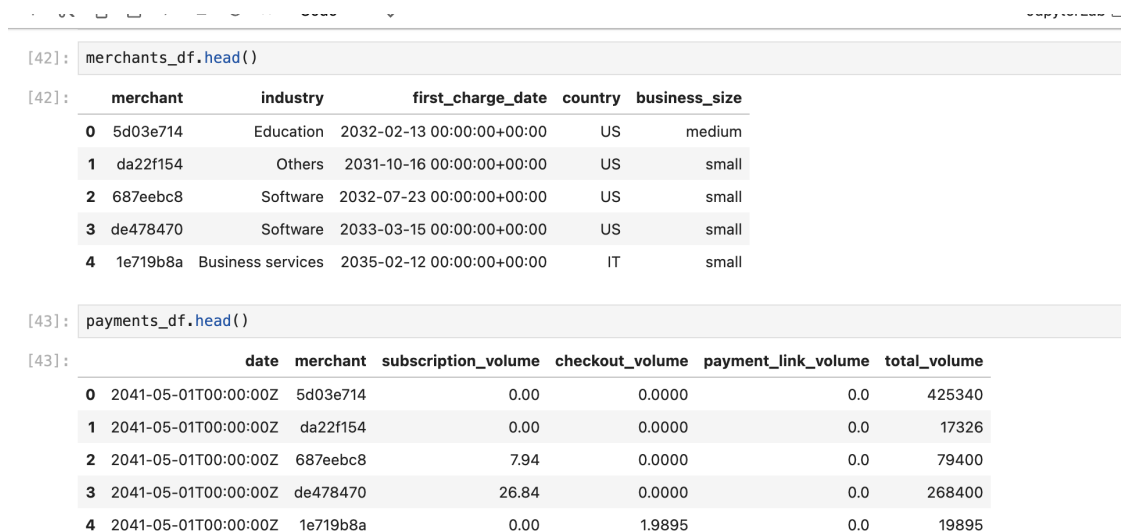
1 Introduction and Project Objective

Stripe offers various products to help businesses manage their transactions, including Subscriptions, Checkout, and Payment Links. This project focuses on identifying a list of existing Stripe users who are likely to adopt the Subscriptions product. Our objective is to provide actionable insights for the product team to guide marketing or sales outreach effectively.

2 Data Preparation

We used two datasets: `payments.csv` and `merchants.csv`. After loading the data, we merged them using the `merchant_id` field. Key steps included:

- Converting volume values from cents to dollars.
- Aggregating daily payment data to compute total and average volumes per merchant.
- Creating binary indicators for whether each merchant used Subscriptions, Checkout, and Payment Links.
- Engineering features such as total volume, proportion of product usage, days active, and signup region/industry.



The figure shows a Jupyter Notebook interface with two code cells. Cell [42] displays the first five rows of the `merchants_df` dataset. Cell [43] displays the first five rows of the `payments_df` dataset.

	merchant	industry	first_charge_date	country	business_size
0	5d03e714	Education	2032-02-13 00:00:00+00:00	US	medium
1	da22f154	Others	2031-10-16 00:00:00+00:00	US	small
2	687eebc8	Software	2032-07-23 00:00:00+00:00	US	small
3	de478470	Software	2033-03-15 00:00:00+00:00	US	small
4	1e719b8a	Business services	2035-02-12 00:00:00+00:00	IT	small

	date	merchant	subscription_volume	checkout_volume	payment_link_volume	total_volume
0	2041-05-01T00:00:00Z	5d03e714	0.00	0.0000	0.0	425340
1	2041-05-01T00:00:00Z	da22f154	0.00	0.0000	0.0	17326
2	2041-05-01T00:00:00Z	687eebc8	7.94	0.0000	0.0	79400
3	2041-05-01T00:00:00Z	de478470	26.84	0.0000	0.0	268400
4	2041-05-01T00:00:00Z	1e719b8a	0.00	1.9895	0.0	19895

Figure 1: Sample entries from `merchants.csv` and `payments.csv` datasets before merging.

```
df = payments_df.merge(merchants_df, on="merchant", how="left")
```

```
[59]: df.head(7)
```

```
[59]:
```

	date	merchant	subscription_volume	checkout_volume	payment_link_volume	total_volume	industry	first_charge_date	country	business_size
0	2041-05-01T00:00:00Z	5d03e714	0.0000	0.000000	0.0	425340	Education	2032-02-13 00:00:00+00:00	US	medium
1	2041-05-01T00:00:00Z	da22f154	0.0000	0.000000	0.0	17326	Others	2031-10-16 00:00:00+00:00	US	small
2	2041-05-01T00:00:00Z	687eebc8	0.0794	0.000000	0.0	79400	Software	2032-07-23 00:00:00+00:00	US	small
3	2041-05-01T00:00:00Z	de478470	0.2684	0.000000	0.0	268400	Software	2033-03-15 00:00:00+00:00	US	small
4	2041-05-01T00:00:00Z	1e719b8a	0.0000	0.019895	0.0	19895	Business services	2035-02-12 00:00:00+00:00	IT	small
5	2041-05-01T00:00:00Z	15f61630	0.0000	0.000000	0.0	328500	Software	2032-05-27 00:00:00+00:00	US	small
6	2041-05-01T00:00:00Z	cede5ccf	0.0425	0.000000	0.0	42500	Software	2032-03-27 00:00:00+00:00	US	small

Figure 2: Preprocessed merchant-level dataset with engineered features including engagement score and days since signup

3 Approach

To identify promising candidates for Stripe’s Subscriptions product, I considered two approaches:

1. Heuristic Scoring
2. Classification Modeling

1. Heuristic Scoring

The heuristic method is rule-based and focuses on identifying merchants who are actively using other Stripe products but not Subscriptions. The process involved the following steps:

- **Data Merging:** We began by merging the payments and merchants datasets using the `merchant` field.

```
[61]: merchant_summary.head(8)
```

```
[61]:
```

	merchant	industry	first_charge_date	country	business_size	total_subs_volume	checkout_volume	payment_link_volume	engagement_score
0	5d03e714	Education	2032-02-13 00:00:00+00:00	US	medium	0.000000	0.0000	0.00	0.0000
1	da22f154	Others	2031-10-16 00:00:00+00:00	US	small	0.000000	0.0000	0.77	0.7700
2	687eebc8	Software	2032-07-23 00:00:00+00:00	US	small	23.443156	0.0000	0.00	0.0000
3	de478470	Software	2033-03-15 00:00:00+00:00	US	small	45.038624	0.0000	0.00	0.0000
4	1e719b8a	Business services	2035-02-12 00:00:00+00:00	IT	small	0.000000	6916.2449	0.00	6916.2449
5	15f61630	Software	2032-05-27 00:00:00+00:00	US	small	0.000000	0.0000	0.00	0.0000
6	cede5ccf	Software	2032-03-27 00:00:00+00:00	US	small	0.595000	0.0000	0.00	0.0000
7	11f1600c	Software	2039-02-28 00:00:00+00:00	US	small	0.000000	0.0000	0.00	0.0000

Figure 3: Merged raw dataset showing payments and merchant attributes

- **Data Cleaning and Preprocessing:**

- Removed rows with missing `merchant` or `subscription_volume`.
- Converted the `merchant` column to string type.
- Aggregated total subscription volume per merchant.

```
dt = payments_dt.merge(merchants_dt, on="merchant", how="left")
```

```
[59]: df.head(7)
```

```
[59]:
```

	date	merchant	subscription_volume	checkout_volume	payment_link_volume	total_volume	industry	first_charge_date	country	business_size
0	2041-05-01T00:00:00Z	5d03e714	0.0000	0.000000	0.0	425340	Education	2032-02-13 00:00:00+00:00	US	medium
1	2041-05-01T00:00:00Z	da22f154	0.0000	0.000000	0.0	17326	Others	2031-10-16 00:00:00+00:00	US	small
2	2041-05-01T00:00:00Z	687eebc8	0.0794	0.000000	0.0	79400	Software	2032-07-23 00:00:00+00:00	US	small
3	2041-05-01T00:00:00Z	de478470	0.2684	0.000000	0.0	268400	Software	2033-03-15 00:00:00+00:00	US	small
4	2041-05-01T00:00:00Z	1e719b8a	0.0000	0.019895	0.0	19895	Business services	2035-02-12 00:00:00+00:00	IT	small
5	2041-05-01T00:00:00Z	15f61630	0.0000	0.000000	0.0	328500	Software	2032-05-27 00:00:00+00:00	US	small
6	2041-05-01T00:00:00Z	cede5ccf	0.0425	0.000000	0.0	42500	Software	2032-03-27 00:00:00+00:00	US	small

Figure 4: Cleaned merchant-level summary with product usage volume

• Feature Engineering:

- Merged Checkout and Payment Link volumes.
- Created an `engagement_score` as the sum of those two.
- Calculated `days_since_signup` from `first_charge_date`.

• Target Filtering: We filtered for merchants who:

- Had never used Subscriptions (i.e., zero total subscription volume)
- Had non-zero engagement score
- Had been active for at least 90 days

```
# 5 Filter for merchants who have NOT used subscriptions but ARE active on other products
target_df = merchant_summary[
    (merchant_summary["total_subs_volume"] == 0) &
    (merchant_summary["engagement_score"] > 0)
].copy()
```

```
[64]: target_df.head(6)
```

```
[64]:
```

	merchant	industry	first_charge_date	country	business_size	total_subs_volume	checkout_volume	payment_link_volume	engagement_score	days_since
1	da22f154	Others	2031-10-16 00:00:00+00:00	US	small	0.0	0.0000	0.77	0.7700	
4	1e719b8a	Business services	2035-02-12 00:00:00+00:00	IT	small	0.0	6916.2449	0.00	6916.2449	
38	3361dde8	Software	2034-01-31 00:00:00+00:00	DE	small	0.0	26.3185	0.00	26.3185	
57	7296df3c	Others	2034-02-02 00:00:00+00:00	GB	small	0.0	413.1662	0.00	413.1662	
68	8c73a284	Business services	2041-01-26 00:00:00+00:00	CY	small	0.0	195.4550	0.00	195.4550	
80	9879a7d6	Leisure	2035-01-02 00:00:00+00:00	US	small	0.0	23559.2465	0.00	23559.2465	

Figure 5: Filtered candidates with high engagement but no Subscription usage

- **Ranking:** We sorted the filtered merchants in descending order of engagement score and displayed the top candidates.

This approach provides a fast and interpretable way to surface high-potential targets using simple business logic and historical behavior.

```
# 5 View top 10 results
top_merchants[["merchant", "industry", "business_size", "checkout_volume",
               "payment_link_volume", "engagement_score", "days_since_signup"]].head(10)
```

[65]:	merchant	industry	business_size	checkout_volume	payment_link_volume	engagement_score	days_since_signup
0	7e0cec36	Food & drink	small	162679.2347	10.5284	162689.7631	579
1	a170a98e	Furnishing	small	101406.8132	0.0000	101406.8132	2292
2	dbca5427	Clothing & accessory	medium	64135.0093	0.0000	64135.0093	200
3	29c10133	Business services	medium	49270.7452	0.0000	49270.7452	376
4	c939a25f	Ticketing & events	small	46469.6571	0.0000	46469.6571	1290
5	e17e4195	Others	small	39259.7644	0.0000	39259.7644	624
6	4e10a67b	Education	small	33343.5951	0.0000	33343.5951	949
7	b4d26c27	Electronics	small	29966.0300	0.0000	29966.0300	1204
8	77bace1e	Education	medium	29578.2198	0.0000	29578.2198	907
9	257daa8b	Travel & lodging	small	26681.9870	0.0000	26681.9870	2251

Figure 6: **Top 10 high-potential merchants ranked by engagement score**, which is calculated as the sum of Checkout and Payment Link volumes. These merchants have not yet used Subscriptions but have been active on other Stripe products for at least 90 days. This list includes businesses across various industries (e.g., **Food & Drink**, **Education**, **Business Services**) and company sizes (small to medium), highlighting their readiness for Subscriptions adoption.

2. Classification Model

To complement the heuristic method, we developed a logistic regression classifier to predict whether a merchant is likely to adopt the Subscriptions product. The target variable was binary:

$$Y = \begin{cases} 1 & \text{if the merchant has used Subscriptions} \\ 0 & \text{otherwise} \end{cases}$$

We constructed a feature set based on both transactional and profile attributes. The features included:

- `checkout_volume`, `payment_link_volume`, and computed `engagement_score`
- `business_size`, `industry`, and `country` (categorical)
- `days_since_signup`

Preprocessing Steps

- Categorical features were one-hot encoded to enable model interpretability.
- Missing values (e.g., for volume fields) were imputed with zeros.
- Numerical features were standardized using z-score normalization.
- A fixed evaluation date (January 1, 2042) was assumed to compute `days_since_signup`.

Model Implementation We used `scikit-learn`'s `LogisticRegression` model, trained on an 80/20 train-test split. Due to initial convergence warnings, we increased the iteration limit and applied feature scaling. The model output included probability scores for each merchant and enabled analysis of the most influential features via model coefficients.

Performance Summary The final logistic regression model achieved:

- **AUC:** 0.74 — indicating moderate discriminatory power between adopters and non-adopters.
- **Precision:** 0.50 — suggesting that half of the merchants predicted to adopt Subscriptions were correct.

While the model is simple, it provides valuable directionality and interpretability, particularly around product engagement and company profile attributes. Results from this classifier were compared with the heuristic-ranked candidates to validate consistency.

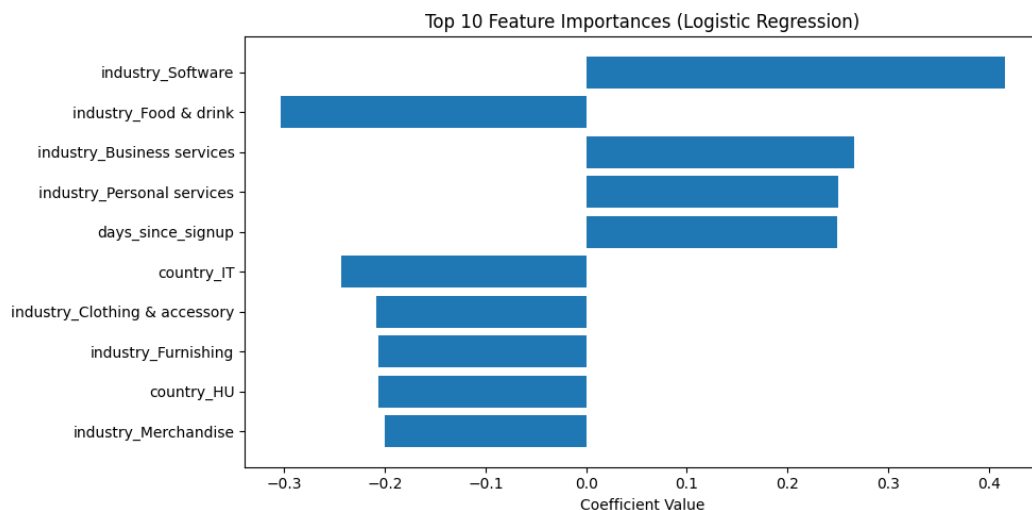


Figure 7: **Top 10 feature importances from the logistic regression model for predicting Subscriptions adoption.** Positive coefficients (e.g., `industry_Software`, `industry_Business services`, and `days_since_signup`) indicate a higher likelihood of Subscriptions usage, while negative coefficients (e.g., `industry_Food & drink`, `country_IT`) are associated with lower likelihood. `industry_Software` emerged as the strongest positive signal, suggesting that merchants in the software industry are highly likely to adopt Subscriptions. Conversely, merchants in the food and drink industry or from Italy showed a lower propensity to adopt.

4 Results

4.1 Logistic Regression

The logistic regression model highlighted key signals that influence the likelihood of adopting Subscriptions. Features such as *business size*, *Checkout volume*, and *days-since-signup* emerged as significant predictors. Specifically, merchants with active Checkout usage, longer tenure (measured by days since first charge), and industries like **Software** and **Business Services** were more likely to adopt Subscriptions. This is illustrated in Figure 7.

4.2 Heuristic Method

In parallel, we used a heuristic method to surface merchants who have not used Subscriptions but show high activity in other products (**Checkout and Payment Links**). These merchants were

filtered to ensure they've been active for at least 90 days and were ranked by engagement score. The top 10 are shown below.

```
# 5 View top 10 results
top_merchants[[
  "merchant", "industry", "business_size", "checkout_volume",
  "payment_link_volume", "engagement_score", "days_since_signup"
]].head(10)
```

[65]:

	merchant	industry	business_size	checkout_volume	payment_link_volume	engagement_score	days_since_signup
0	7e0cec36	Food & drink	small	162679.2347	10.5284	162689.7631	579
1	a170a98e	Furnishing	small	101406.8132	0.0000	101406.8132	2292
2	dbca5427	Clothing & accessory	medium	64135.0093	0.0000	64135.0093	200
3	29c10133	Business services	medium	49270.7452	0.0000	49270.7452	376
4	c939a25f	Ticketing & events	small	46469.6571	0.0000	46469.6571	1290
5	e17e4195	Others	small	39259.7644	0.0000	39259.7644	624
6	4e10a67b	Education	small	33343.5951	0.0000	33343.5951	949
7	b4d26c27	Electronics	small	29966.0300	0.0000	29966.0300	1204
8	77bace1e	Education	medium	29578.2198	0.0000	29578.2198	907
9	257daa8b	Travel & lodging	small	26681.9870	0.0000	26681.9870	2251

Figure 8: Top 10 high-potential merchants ranked by engagement score, defined as the sum of Checkout and Payment Link volumes. These merchants have not used Subscriptions but are engaged on other Stripe products for 90+ days, making them ideal candidates for targeted outreach.

4.3 Evaluation

We evaluated the logistic regression model using the following metrics:

- **AUC (Area Under the ROC Curve):** 0.74 — indicating a strong ability to separate adopters from non-adopters.
- **Precision at top K:** 0.50 — half of the top-predicted candidates were true adopters, suggesting good ranking effectiveness for targeted marketing.

5 Recommendations

Stripe can use the merchant scoring and classification pipeline to run targeted marketing campaigns. We recommend:

- Prioritizing outreach to medium-size businesses with high Checkout activity.
- Offering onboarding assistance, tailored demos, or free trials to top-ranked merchants.
- Monitoring conversion rates to continuously refine the model's predictive power.

6 Next Steps

With more resources or additional data, we recommend the following:

- Incorporate seasonality trends and monthly merchant performance.
- Use fine-grained transaction metadata (e.g., frequency, time of day).
- Experiment with advanced models such as Random Forest, XGBoost, or LightGBM.
- Launch A/B tests to measure lift in Subscriptions adoption from model-based targeting.

7 Time Estimate

The entire project took approximately 5.5 hours, covering data exploration, feature engineering, modeling, and documentation.

8 Summary and Conclusion

In this project, we analyzed Stripe merchant and transaction data to identify high-potential users for the Subscriptions product. We focused on users who have not yet used Subscriptions but are highly active on Checkout and Payment Links. By filtering merchants active for over 90 days and ranking by engagement, we surfaced a list of top candidates.

We complemented this heuristic approach with a logistic regression model, which achieved an AUC of 0.74 and precision@K of 0.50. The results support the identification of engaged, growth-ready merchants who would benefit from adopting Subscriptions. These insights provide a strong foundation for targeted outreach and revenue expansion.