**Relax Inc Take Home challenge Overall Summary**

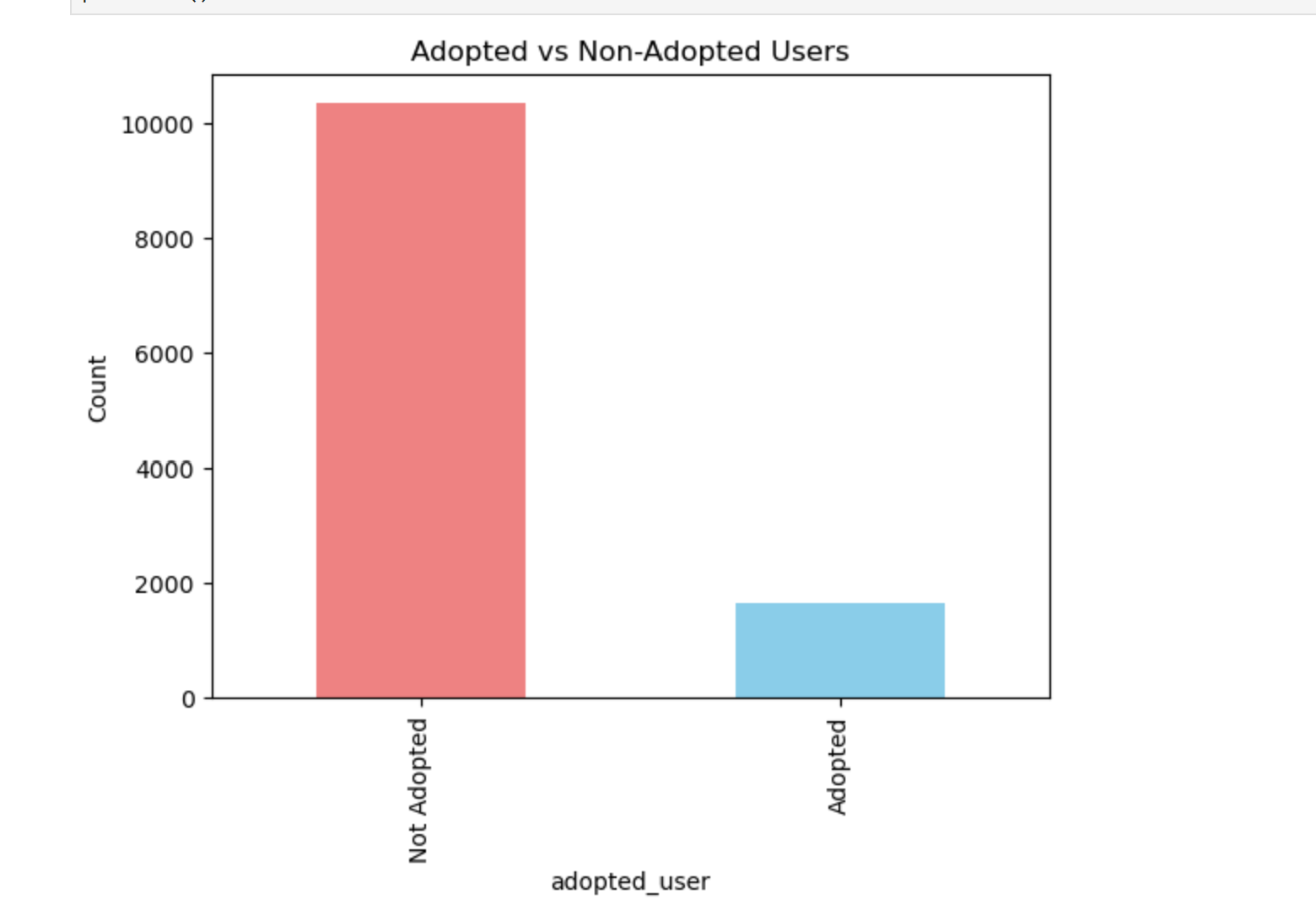
User Adoption Analysis and Prediction

1. **Dataset** Overview: We worked with two datasets:

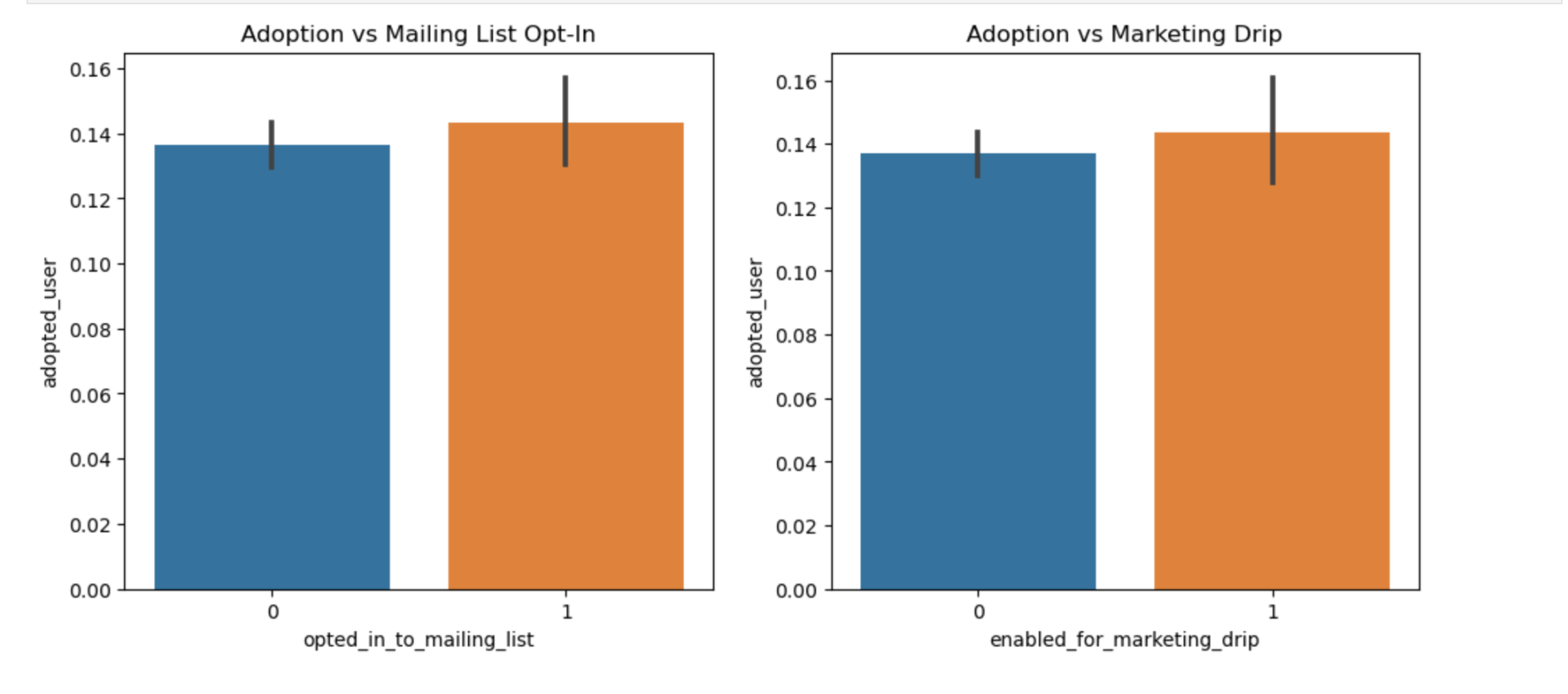
* **users.csv**: contains user-level attributes (signup date, signup source, organization ID, invited flag, mailing list opt-in, etc.).
* **engagement.csv**: contains user activity logs (logins, timestamps).

The target variable **adopted\_user** was created by labeling users as "adopted" if they logged in on 3 separate days within any 7-day window. Out of 12,000 users, 1,656 were adopted (~13.8%), showing strong class imbalance.

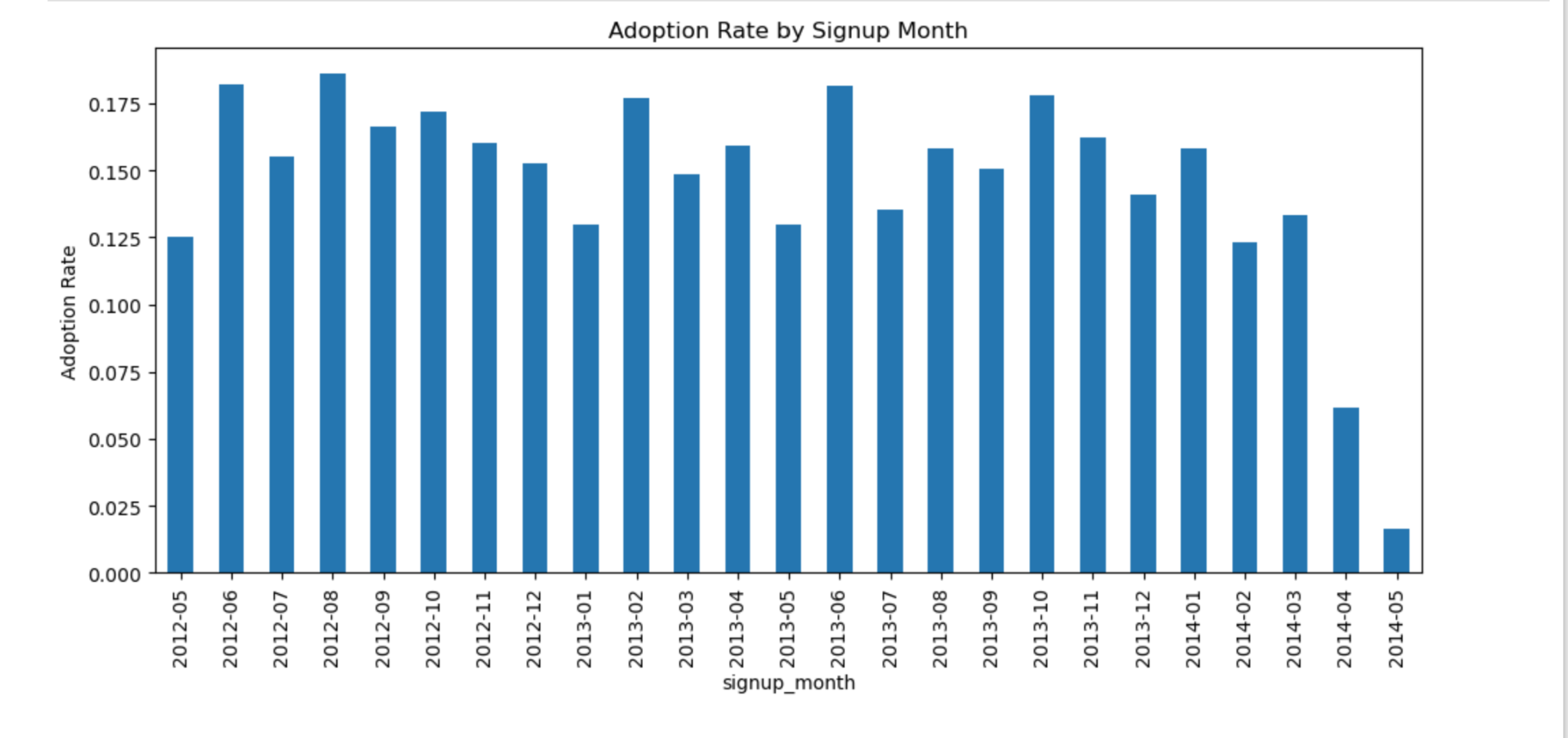
1. **Exploratory Data Analysis(EDA):**



The distribution shows a significant class imbalance: around 86% of users are non-adopted, while only about 14% are adopted. This imbalance makes adoption prediction more challenging.



Adoption rates are only slightly higher for users who opted into the mailing list or marketing drip (~14.3% vs ~13.7%). The effect is minimal, suggesting that marketing participation does not strongly influence adoption.

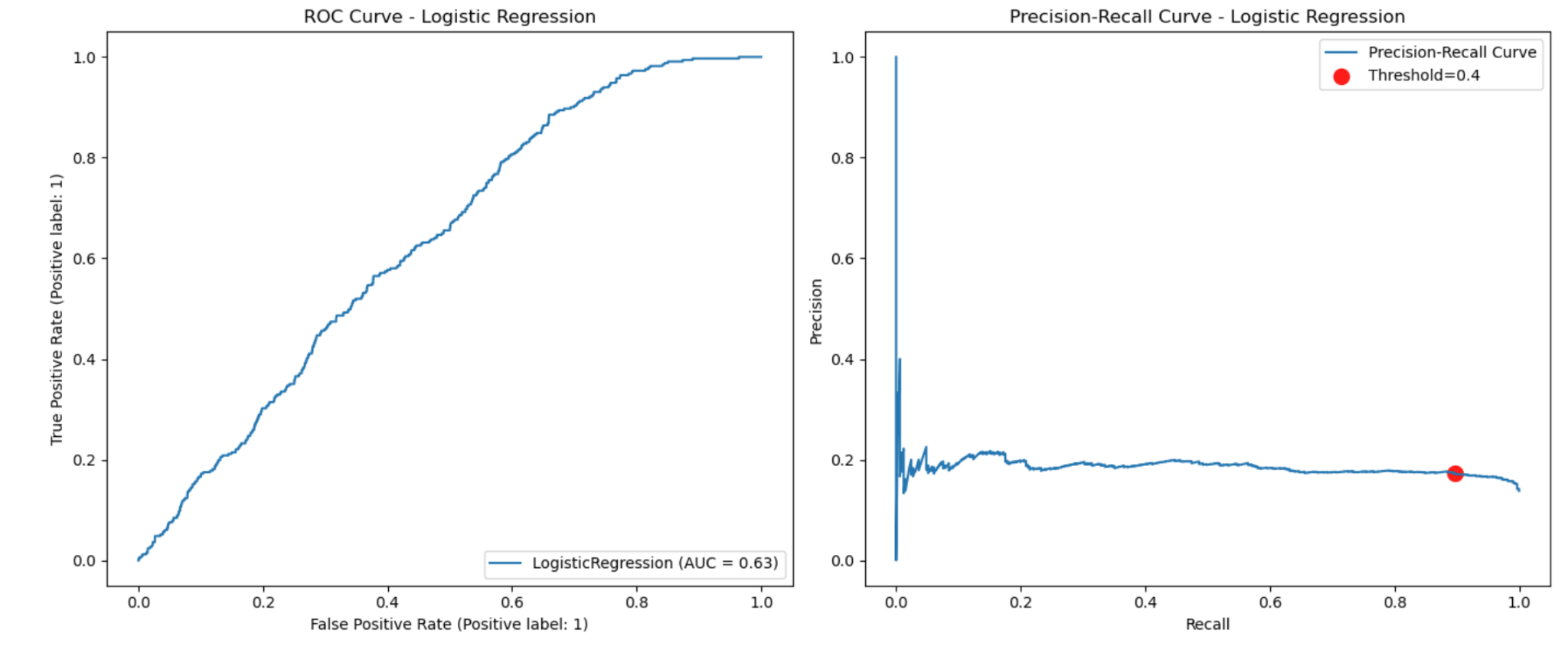


Adoption rates remained steady between 2012–2013 (around 15–19%) but declined sharply in 2014, dropping below 10% by April–May. This suggests that later cohorts were less engaged despite increasing signups.

1. **Preprocessing**:

* Removed irrelevant identifiers (user\_id, name, email).
* Standardized date fields:  
   Converted creation\_time into **signup month (cohort)**.  
   Transformed last\_session\_creation\_time into datetime format.
* Engineered new features:  
   **Invited flag** → whether a user was invited by someone.  
   **Organization size** → number of users within the same org\_id.
* Cleaned binary marketing flags (opted\_in\_to\_mailing\_list, enabled\_for\_marketing\_drip) and converted them into integers.

1. **Modelling**:



* **Logistic Regression:**
  + ROC-AUC = **0.632** (moderate performance).
  + High recall (**0.713**) but low precision (**0.175**).
  + Key positive drivers: early cohorts, invitations, Google Auth.
  + Negative drivers: late 2014 cohorts, ORG\_INVITE, Personal Projects.
* **Random Forest:**
  + ROC-AUC = **0.629**, close to logistic regression.
  + Recall weaker (**0.375**) but slightly better precision (**0.204**).
  + Less interpretable than logistic regression.
* **Threshold Tuning (Logistic):**
  + Optimal threshold ≈ **0.4**.
  + Recall increased to **0.897**, but precision stayed low.
  + Good for maximizing adopter detection, but at the cost of more false positives.

1. **Key Takeaways:**

* **Adoption is strongly influenced by signup cohort and signup source.**
  + Early adopters (2012–2013) had higher engagement.
  + Users signing up through **Google Auth** or via an **invitation** were more likely to adopt.
  + Late 2014 signups were much less likely to adopt.
* **Marketing features (mailing list, drip) had negligible impact** on adoption.
* **Modeling insights:**
  + Logistic Regression outperformed Random Forest and provided interpretability.
  + Threshold tuning improved recall significantly (0.897 at threshold 0.4), but precision remained low.
  + Adoption prediction remains challenging with available features, as precision is consistently weak.

1. **Business Recommendations:**

* Focus on **high-quality signup channels** (Google Auth, invited users).
* Investigate why **late 2014 cohorts** show poor adoption.
* Reconsider marketing strategy since opt-in/drip features show little effect.
* Use the tuned logistic regression (threshold = 0.4) for recall-driven use cases where catching adopters matters more than avoiding false positives.

**Conclusion**: Adoption is best explained by **signup cohort and signup source**, with strong early adopters and invited/Google-authenticated users showing higher engagement. Logistic Regression provides actionable insights, though predictive performance is modest due to class imbalance and limited features. Future work should expand feature engineering and explore boosting models for improved adoption prediction.