Analysis of Customer Data

Insights into Engagement and Segmentation
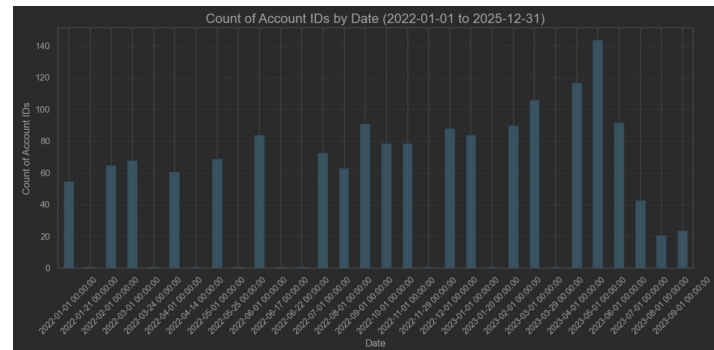By: Shaked Markovich, Nov 2024

EDA

NaN Handling:

- Fill missing values in numeric columns with 0.
- Fill missing values in categorical columns with the most frequent value.

Data Cleaning:

- Remove accounts with low activity and old account records.

EDA

A strong starting point would involve comparing products (P1, P2) across various features (e.g., impressions, engagement) and metrics (such as count, sum, and average). Later, I will demonstrate the integration of features.

I will present a sample of figures here, while the full distribution can be explored in the notebook.

EDA: BASIC

Despite P1 being more frequent, its average and sum of number of technical and support sessions are lower compared to P2.

Also, despite P1 being more frequent, its average impressions are lower compared to P2.
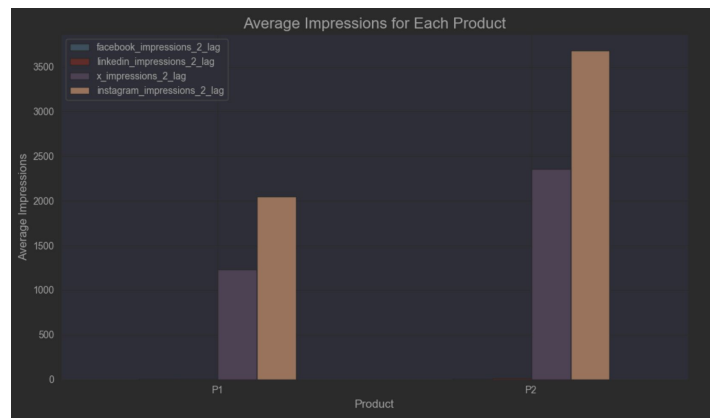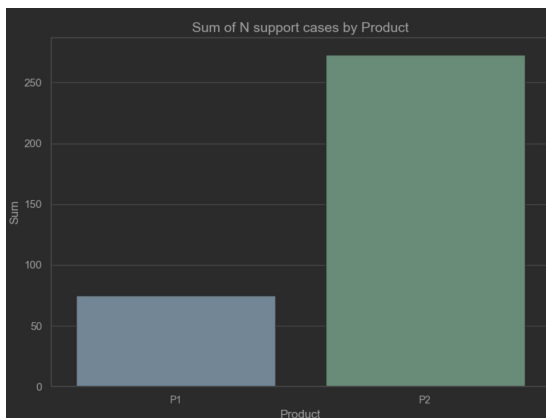
EDA: Territory Analysis

First, let's gain an understanding of the feature by examining the count distribution across products and territories. Afterward, I'll compare engagement metrics and impressions across the different territories (T1, T2, T3).
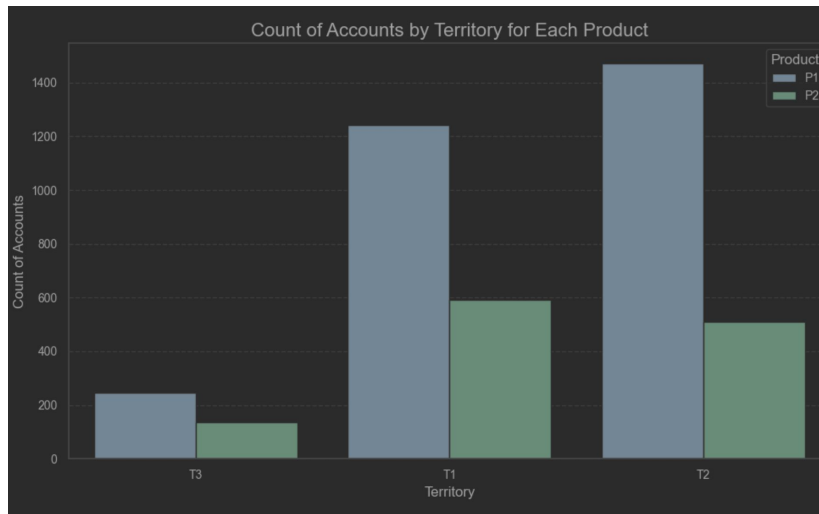
EDA: Territory Analysis - Engagement Metrics

T3 within the P1 tier exhibits a remarkably high average for n_users_1_lag, while T1 demonstrates a high average for n_technical_sessions across both

Average n_users_1_lag by Territory and Product

Average n_technical_sessions by Territory and Product
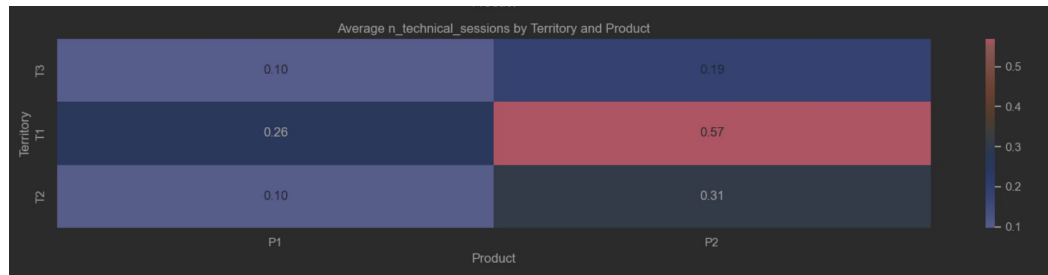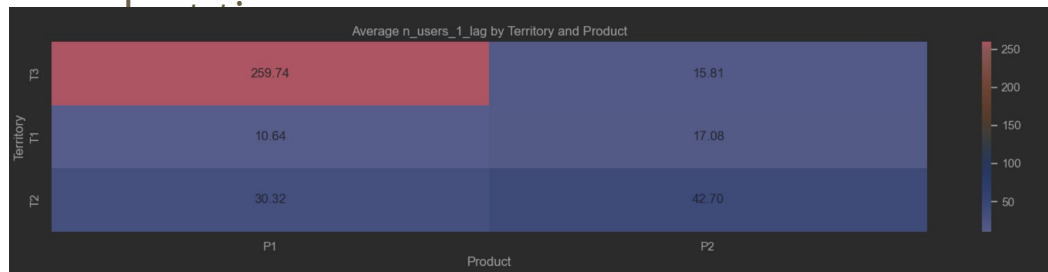
EDA: Territory Analysis - Impressions Metrics

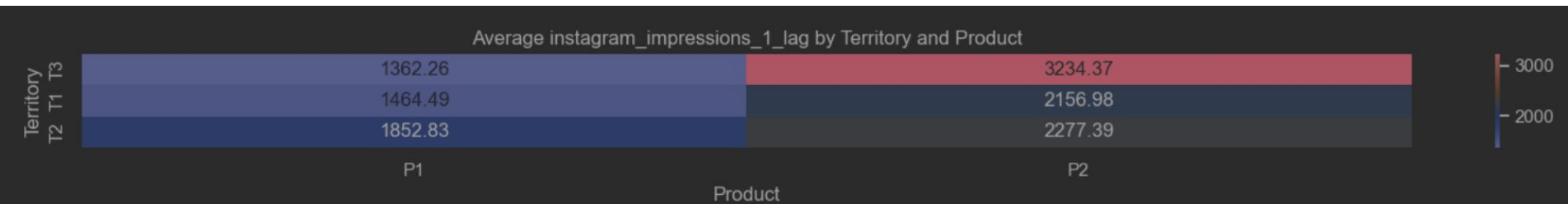T1 in the P1 tier has the highest average Facebook impressions, while T3 in the P2 tier leads with the highest average Instagram impressions.

Average facebook_impressions_1_lag by Territory and Product

Territory	P1	P2
T3	1.57	4.30
T1	10.87	4.02
T2	3.69	7.99

Average instagram_impressions_1_lag by Territory and Product

Territory	P1	P2
T3	1362.26	3234.37
T1	1464.49	2156.98
T2	1852.83	2277.39
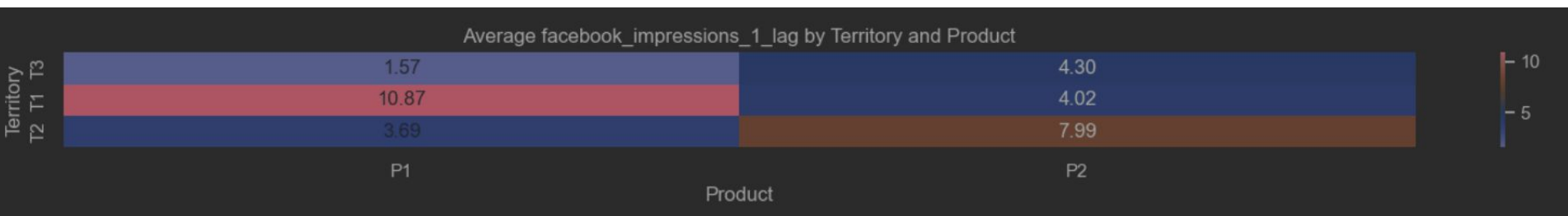
EDA: Territory Analysis - Impressions Metrics

There is a temporal variation in X impressions: T3 has the highest average impressions in lag 1, but in lag 2, the highest shifts to T2.

Average x_impressions_1_lag by Territory and Product

Territory	P1	P2
T3	855.47	1318.07
T1	903.14	1290.82
T2	1062.97	1204.20

Average x_impressions_2_lag by Territory and Product

Territory	P1	P2
T3	600.10	1015.91
T1	877.15	1290.42
T2	1047.73	2489.44

EDA: Industry Trends

P1 is used most by business services, manufacturing and software

Mode to do: product/industry/impressions+engagement metrics as heatmap

Heatmap of Product vs Industry Group Counts

| Product	Agriculture	Business Services	Construction	Consumer Services	Education	Energy, Utilities & Waste	Finance	Government	Healthcare Services	Holding Companies & Conglomerates	Hospitality	Hospitals & Physicians Clinics	Insurance	Law Firms & Legal Services	Manufacturing	Media & Internet	Minerals & Mining	Organizations	Real Estate	Retail	Software	Telecommunications	Transportation	unknown
P1	3	512	23	68	51	45	257	96	9	19	64	20	89	7	543	74	3	23	15	107	684	116	64	68
P2	0	165	6	21	31	30	217	78	2	8	28	12	77	2	219	13	0	8	0	31	202	45	31	11

EDA: Employee Range Influence

On average, smaller companies with P2 tend to have more support cases.
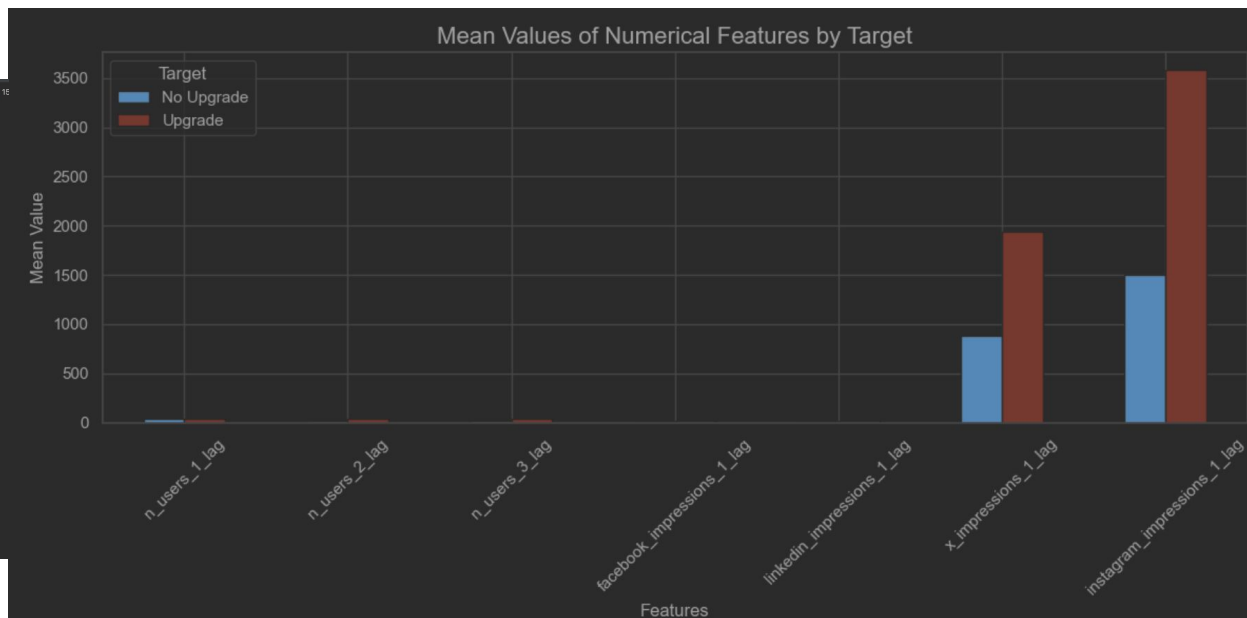In addition, larger companies with P1 generate significantly higher Facebook impressions.
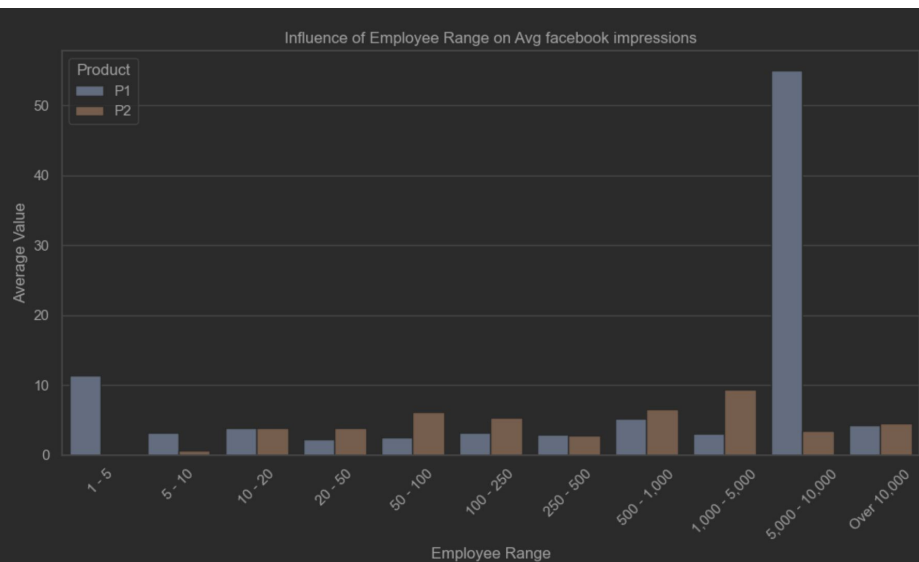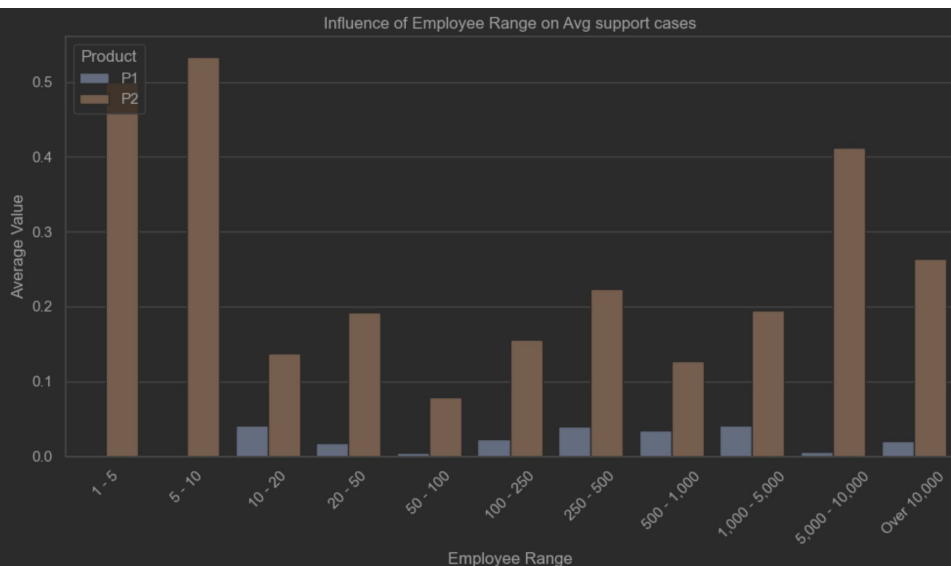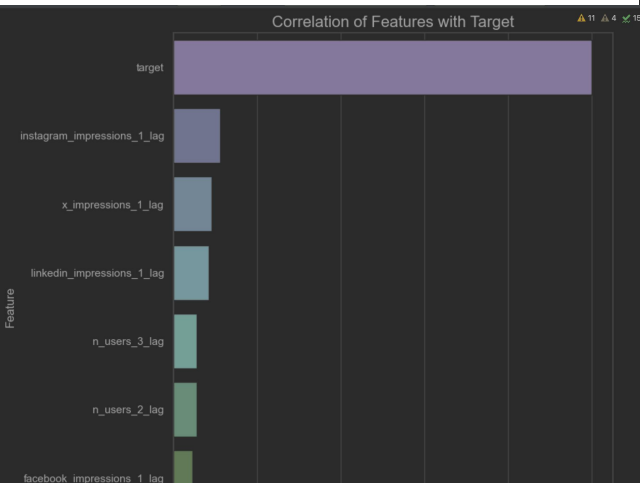
EDA: Analysis of the target Column

Data is biased.
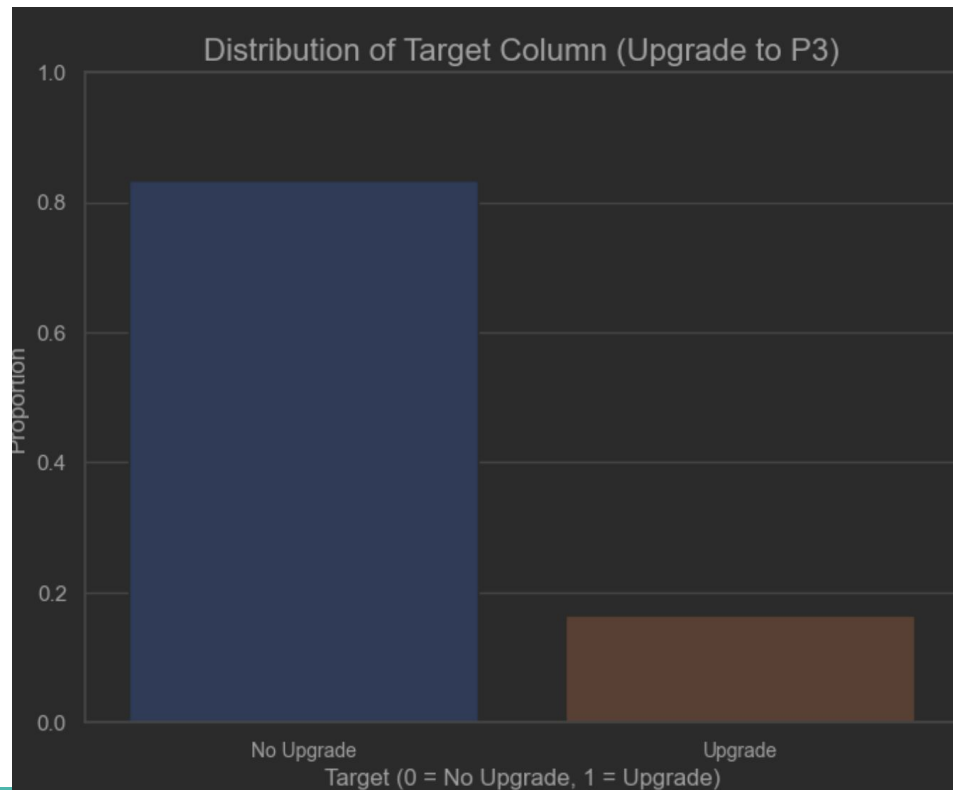
EDA: Relationship Between target and Key Features

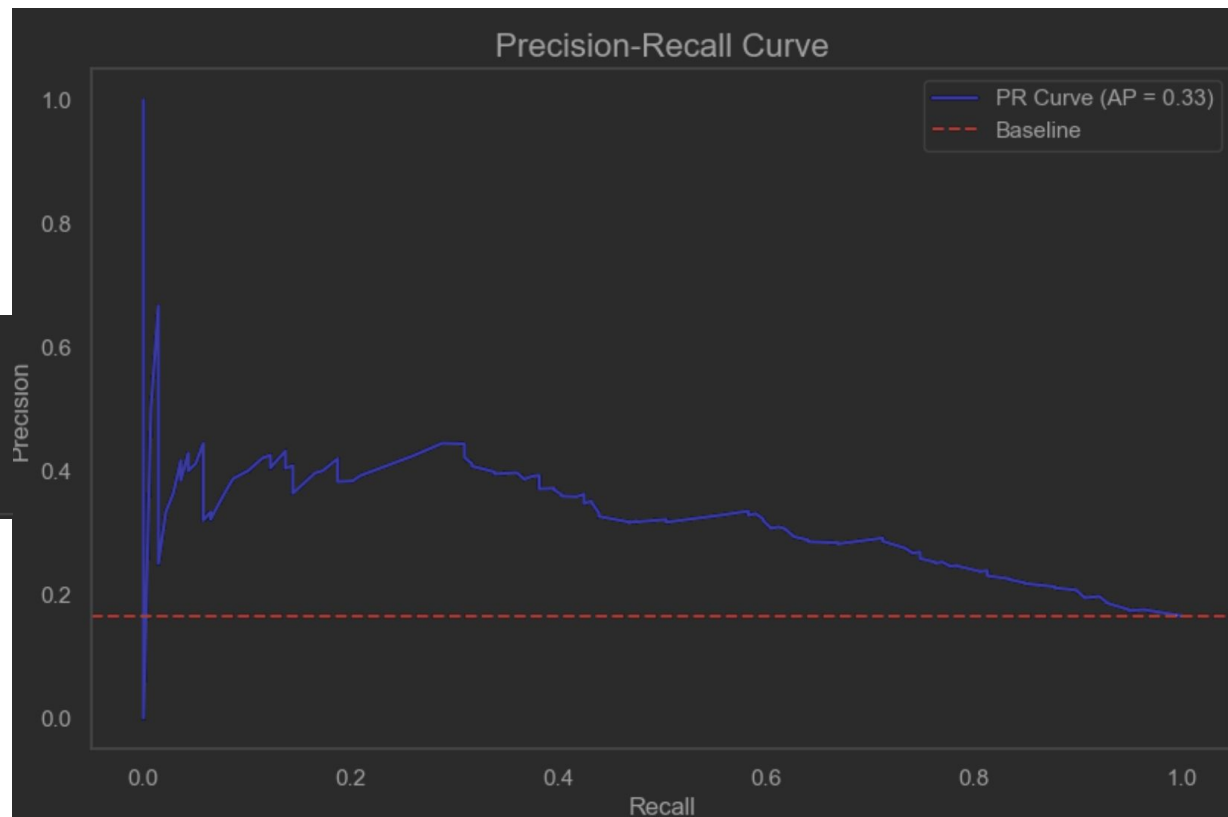Looks like upgraded accounts are accounts with high x and instagram impressions.

Model Results

Precision-Recall AUC: 0.3208
F1-Score: 0.1899
Lift (Top 20%): 2.1223

Precision-Recall Curve

Metrics Discussion

Metric	Pros	Cons
Accuracy	- Simple to calculate and understand.	- Misleading for imbalanced datasets (high accuracy can be achieved by predicting the majority class). - Doesn't focus on the minority class (e.g., rare events).
ROC AUC	- Provides an overall view of model performance, balancing true positive and false positive rates. - Not affected by imbalanced classes as severely as accuracy.	- Can be misleading for highly imbalanced datasets because the false positive rate may dominate. - Doesn't directly focus on performance for the minority class.
PR AUC	- Focuses on the positive class (rare events), providing a better understanding of model performance for imbalanced data. - Measures both precision and recall.	- Can be less intuitive for non-experts. - Less informative for balanced datasets compared to accuracy or ROC AUC.
F1-score	- Balances precision and recall, giving a single value that accounts for both false positives and false negatives. - Especially useful for imbalanced datasets	- Doesn't consider true negatives, so may not fully reflect model performance in certain contexts. - May not be as interpretable in some business scenarios as accuracy

Suggested Action Items for DS team

- **Optimize the Model for Deployment**:
  - Fine-tune hyperparameters of the model (e.g., learning rate, max depth) to further improve Precision-Recall AUC and reduce Log Loss.
  - Investigate additional advanced algorithms (e.g., CatBoost, LightGBM) to benchmark performance.
- **Explore Additional Features**:
  - Engineer interaction terms between impressions, user engagement, and support cases.
- **Conduct Feature Importance Analysis**:
  - Use SHAP values to understand key drivers of customer upgrades to P3 and guide strategic decisions.
- **Handle Imbalanced Data**:
  - Use techniques like SMOTE or class-weighted loss functions to further improve performance on minority class (upgrades to P3).

Suggested Action Items for Business Stakeholders

Target Engagement Strategies:

- **P1 Tier**: Since **P1** users (especially in larger companies) have higher **Facebook impressions**, consider prioritizing **Facebook-based campaigns** or increasing engagement efforts on this platform for **larger clients**.
- **P2 Tier**: For **P2**, which shows higher engagement for **smaller companies** and **Instagram impressions**, explore tailored campaigns or features that appeal more to these businesses, leveraging **Instagram** as a primary channel.

Focus on Underrepresented Territories:

- **T1 and T3**: Given that T1 in **P1** has the highest average **Facebook impressions** and T3 in **P2** has the highest **Instagram impressions**, assess whether marketing efforts can be optimized for these territories by adjusting content to the platform's strength.
- Investigate why certain territories (like T3 in **P1**) show remarkably high **user engagement** (e.g., **n_users_1_lag**) and explore the underlying factors for scalability.

Suggested Action Items for Business Stakeholders

Enhance Support Services:

- Smaller **P2 companies** tend to have more **support cases**, indicating a need for better support structures, such as increased training or resources, for **smaller businesses** using **P2**.

Industry-Specific Strategies:

- Since **P1** is most used in industries like **business services, manufacturing,** and **software**, consider offering industry-specific packages or features that align with the needs of these sectors.