Business Analyst Challenge - Influur

Federico Wolff

1. We just launched the new version of the app to market. As you can probably imagine, data is growing and changing very fast. Hence, making constant monitoring of the engagement with the app a key business need.
What would be the key performance indicators you would come up with as the most important to monitor engagement with the app?
How often would you suggest such indicators must be monitored?

I split the KPIs into General and Financial KPIs, and Engagement KPIs.
General and Financial:

1. **Number of installs:** Number of installs in the store.
2. **New users** (and the relationship with installs, to see if there's a registration issue). Split by:
    a. Brands
    b. Creators
3. **DAU**:
    Split by:
    a. Brands
    b. Creators
4. **LTV**: d0, d1, d3, d7, d14. (over install dates). In LTV we consider users that installed the new app version, and in contrast to ARPDAU, we don't consider users that updated to the new version.
5. **ARPDAU**: Avg Revenue per Daily Active User.
6. **Spend, Revenue and GOM.**
7. **Daily Payers**
8. **CVR and Payment Conversion**: Cohorted and daily
9. **Transaction amount per payer**
10. **Transaction count per payer**
11. **ARPPU**
12. **Transaction number by category buckets.**
13. **ROAS**
14. **Campaigns per Influencer**
15. **Net Promoter Score**

Engagement KPIs: Most of them are Cohort with a daily perspective. It should have a clear uplift if the new version is successful.

1. **Retention Rate**: d0, d1, d3, d7
2. **Session length:** d0, d1, d3, d7
3. **Session count:** d0, d1, d3, d7
4. **DAU** and **WAU**.

Effective KPI monitoring requires adapting the frequency of analysis according to the volume of data and the specific nature of each metric. Ensuring a robust sample size is essential for

meaningful insights across all KPIs. Additionally, certain events, such as in-store payments, may introduce delays in data availability, necessitating vigilance when monitoring these metrics.

In my perspective, a daily monitoring cadence for most KPIs is imperative, particularly to assess the performance of new app versions promptly and to address any potential issues in marketing, product, or data. Maintaining clear communication channels between the Marketing and Data teams is pivotal to understanding data availability timelines and ensuring the completion of data pipelines. Furthermore, close collaboration with the Product team is vital to anticipate KPI changes in new versions, mitigating the risk of analysts working with outdated information, minimizing unproductive time, and preventing the propagation of false alarms.

In my view, it's crucial to consider implementing an A/B test (or version comparison) for the Android platform. Unlike iOS, Android allows for concurrent release of two versions in the app store. This comparison becomes essential to ensure that the same audience experiences both versions simultaneously, facilitating an accurate assessment.

Furthermore, determining the requisite sample size is paramount to effectively compare specific KPIs between the two versions and validate the significance of any observed differences. For instance, if Version A boasts a 1% conversion rate (CVR) while Version B achieves 1.5% CVR, it becomes imperative to ascertain whether the sample sizes for both versions are statistically significant. Equally important is ensuring that these samples are well-distributed across various factors, such as marketing campaigns, creatives, countries, and more, to yield robust insights.

2. Many times a business report or dashboard is needed by different stakeholders, where they may not necessarily have the same interpretation of a specific concept. Let's take for example the concept 'active user': some stakeholders may interpret an active user as one that logs into the app with a certain frequency, while others may not consider a simple log-in as a relevant engagement.
How would you propose a problem resolution strategy with the stakeholders? Which facts would you present to them?

Resolving differing interpretations of a concept like "active user" among stakeholders is essential to ensure that everyone involved is on the same page.

I would initiate the process by establishing a clear definition for the concept "active user" within our specific context. This definition must be documented and readily accessible to all stakeholders.

In cases where some stakeholders believe that this definition may not fully capture user engagement, we can introduce a complementary concept, perhaps named "Engaged Active User." This new KPI would align with the original definition while accommodating those stakeholders' preferences. For instance, "Engaged Active User" could represent users who

spend more than one minute on the platform daily. Additionally, we can track the trend of a new metric "Engaged Active Users per Active Users" to provide a more comprehensive view.

As emphasized from the outset, it's crucial that everyone within the organization is well-informed about these definitions and their implications. This applies to various teams, from Data Engineers responsible for data calculation and integration, including partnerships with user acquisition channels, to our C-level executives who have a vested interest in understanding and monitoring the company's KPIs. Clear communication and alignment on these concepts will enable us to make more informed decisions and drive our business forward effectively.

3. *It is a common practice to have many systems scattered all over: one might be hosting the Influur app, and others might be hosting models needed for daily operations. This usually benefits usability over scalability. Nevertheless, data centralization is crucial for its exploitation. For simplicity, imagine there are 4 systems:*

- *The first system hosts the app. It generates data that is stored in an internal database (ignore the database's architecture for now). Every time influencers and brands interact with a screen, click a button, or open the app, this is stored as an event.*
- *The second system hosts the AI matching engine. Every time a brand asks for a recommended influencer, the system retrieves the best matches from Influur's AI matching engine.*
- *The third system hosts influencers' and brand's information. Here, unrestricted information is hosted. This database contains the name, contact info, relevant Instagram info, followers, etc...*
- *Finally, the fourth and last system hosts all the payment information, this means, all the information related to past services: the brand that paid for the services, the influencer that provided the service, payments, etc...*

All systems share a unique identifier for all of the influencers and brands. Those are the keys that allow data to be joined on other databases.
What should we do to centralize the data in order to display it in charts for KPI monitoring? What would you propose the data governance strategy should be?

I propose the implementation of a centralized data warehousing solution as the linchpin of our data consolidation efforts. This data warehouse serves as the central repository for amalgamating data from our diverse array of sources. Although the data format may not yet align perfectly, our primary objective is to capture data comprehensively from all sources.

To achieve this, we can establish ETL processes within a robust pipeline framework, such as Airflow. These ETL processes will be designed to extract data from each source system, harmonize it into a consistent format, and subsequently load it into the data warehouse. Importantly, these ETL processes will be scheduled to run at regular intervals, ensuring that our data remains current and reflective of real-time operations.

Following the data ingestion process, we can leverage the unique identifiers shared across all our systems as the linchpin for cross-system data linkage within the data warehouse. This strategic approach ensures data cleanliness and accessibility for end-users.

In terms of data table design, it is imperative to consider the anticipated querying patterns. For instance, if users frequently query Daily Aggregated Metrics, it is prudent to optimize the pipeline by running it twice daily, generating an aggregated table. This table can then be queried directly by users, reducing computational overhead and enhancing internal efficiency.

To fortify our data governance strategy, we will diligently assign data ownership responsibilities to specific teams or individuals within our organization. These stewards of data quality and accuracy will work tirelessly to uphold our high standards. Regular data quality checks, which can trigger alarms for anomalies or data quality issues, will be a core part of our approach.

Compliance with data privacy regulations, exemplified by our strict adherence to GDPR in previous endeavors, is non-negotiable. Our commitment extends to metadata management, which ensures suggested values are derived from metadata tables, fostering efficiency and compliance.

Finally, documentation will be a cornerstone of our approach, encompassing data sources, the data pipeline, the Enterprise Data Repository (EDR), and table schemas along with their associated keys. This documentation will be readily accessible and serve as a vital reference point.

In summary, with this strategy our data centralization and governance strategy will not only ensure data reliability and security but also empower efficient chart generation and KPI monitoring. By championing this approach, we lay the groundwork for data-driven decision-making and long-term scalability in our analytics and reporting capabilities.

4. Creatively design charts and tables to best describe relevant data. Generate a set of those key performance indicators you consider that drive the business. Present recommendations based on those indicators that, to the best of your knowledge, might be low or could be improved.

With the data shared, I can share some business insights. 31% of the Communications Sent, finish with a Delivered Card. But only 69% of approved transactions have a delivered card. It is important to understand if this percentage is low because the card hasn't been delivered yet or there's an issue delivering the credit card. Also, the delivery score is 2 points out of 5. Can we read comments with the reason we have a low Avg Delivery Score?
Lastly, 50% of the rejected cards are due to USAGE and MOP. Can we investigate more about Usage and MOP reasons?

I began my analysis by exploring and scrutinizing the dataset within Google Colab. During this initial phase, I made several key data modifications. Specifically, I introduced a "NO REPLY" value in cases where both the Status and Txn fields were null, as stipulated in the test requirements. Furthermore, I delved into the UPDATE values, although I encountered no apparent correlations. Even after ordering the data by ID and Status, I observed instances where the event time exhibited both increases and decreases. It is imperative that we pinpoint the root cause of these discrepancies in the UPDATE data source, as this column holds significant importance in comprehending evolving trends.

Despite the inherent unreliability of UPDATE values, I proceeded to construct a Dashboard that displays trends over time, focusing on date-based analysis. Additionally, it is crucial to represent the UPDATE field as datetime, as this facilitates our ability to identify bottlenecks in the process. This approach enables us to address pertinent questions such as "Which step is experiencing prolonged delays? Is it specific to certain countries or days of the week?"

With the data at hand, I can now offer some valuable business insights. Notably, 31% of Communications Sent culminate in a Delivered Card, whereas only 69% of approved transactions are associated with a delivered card. It is imperative to discern whether this lower percentage arises due to delayed card deliveries or potential issues in the credit card delivery process itself. Additionally, it is worth noting that our delivery score stands at a mere 2 out of 5 points. Can we access comments that shed light on the reasons behind this suboptimal Average Delivery Score?

Lastly, an intriguing finding emerged: 50% of the rejected card cases can be attributed to USAGE and MOP (Method of Payment) issues. It would be prudent to conduct a more in-depth investigation into the root causes of these Usage and MOP-related rejections.