# Internship Report: User Adoption Path Data-Driven Framework

## Student Details

**Student Name:** Parag Jain

**Student ID:** 20213682

# Title of the Internship Project

**User Adoption Path Data-Driven Framework for Identifying Users Similar to Churned Users**

## Host Organization Details

- **Organization Name:** Dreeven Technologies
- **Contact Information:** (514) 532-0155
- **Definition:** Dreeven Technologies is a collaborative project management software aimed at elevating construction industry standards and simplifying workflows.
- **Internship Supervisor:** Marzieh Zare
- **Internship Supervisor's Email:** marzieh.zare@dreeven.com

# Executive Summary

In today's digital landscape, user engagement and retention are critical factors for the success of online platforms. This internship report presents the development and implementation of a data-driven framework at Dreeven Technologies to analyze user adoption paths and predict churn behavior. The report outlines the project objectives, methodology, findings, challenges encountered, and recommendations for future enhancements.

## Table of Contents

# 1. Introduction

In the era of digital platforms, understanding user behavior is paramount for ensuring user satisfaction and business success. This internship project focuses on leveraging data analytics to identify user adoption patterns and predict churn, ultimately aiming to enhance user retention strategies at Dreeven Technologies.

## 2. Objectives of the Internship Project

The primary objectives of this internship project are:

- Analyze user adoption paths to identify patterns associated with churned users.
- Develop predictive models to forecast users at risk of churn based on their behavior.
- Implement proactive intervention strategies to retain at-risk users and improve overall user retention rates.
- Evaluate the effectiveness of the data-driven framework in reducing churn and enhancing user engagement.

## 3. Dataset Description and Pre-processing

The internship project utilized real-time user activity logs from the Dreeven Platform. The dataset underwent extensive pre-processing to clean, structure, and standardize the raw log entries. This involved extracting relevant attributes such as timestamp, action type, user ID, and platform details.

# 4. Exploratory Data Analysis (EDA) Using Data Visualizations

Exploratory Data Analysis (EDA) plays a pivotal role in understanding the underlying patterns and characteristics of the dataset. Throughout this internship project, various data visualizations were employed to explore and gain insights into the clickstream data. User log data was combined with various other data sources to understand the data in more detail. Using the data available, we tried to combine it and create visualizations so that we can understand what could be going on at user level, platform level and company level within Dreeven.

Following is the snapshot of the visualizations created on Tableau :

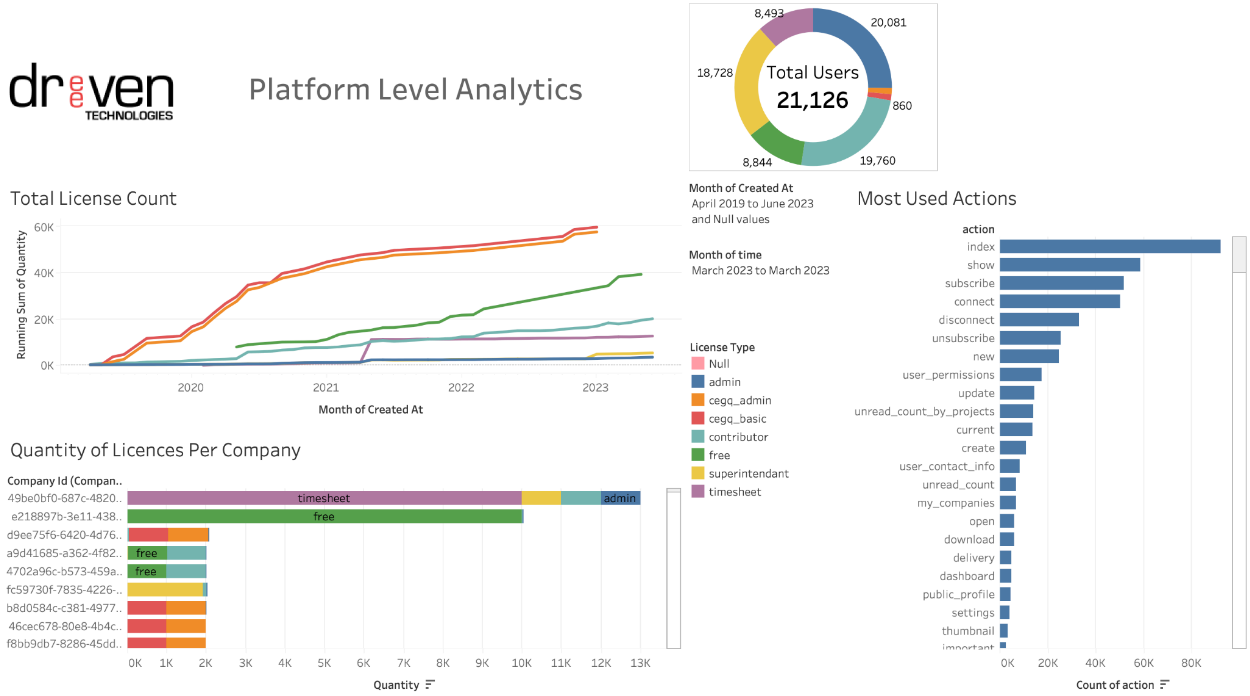


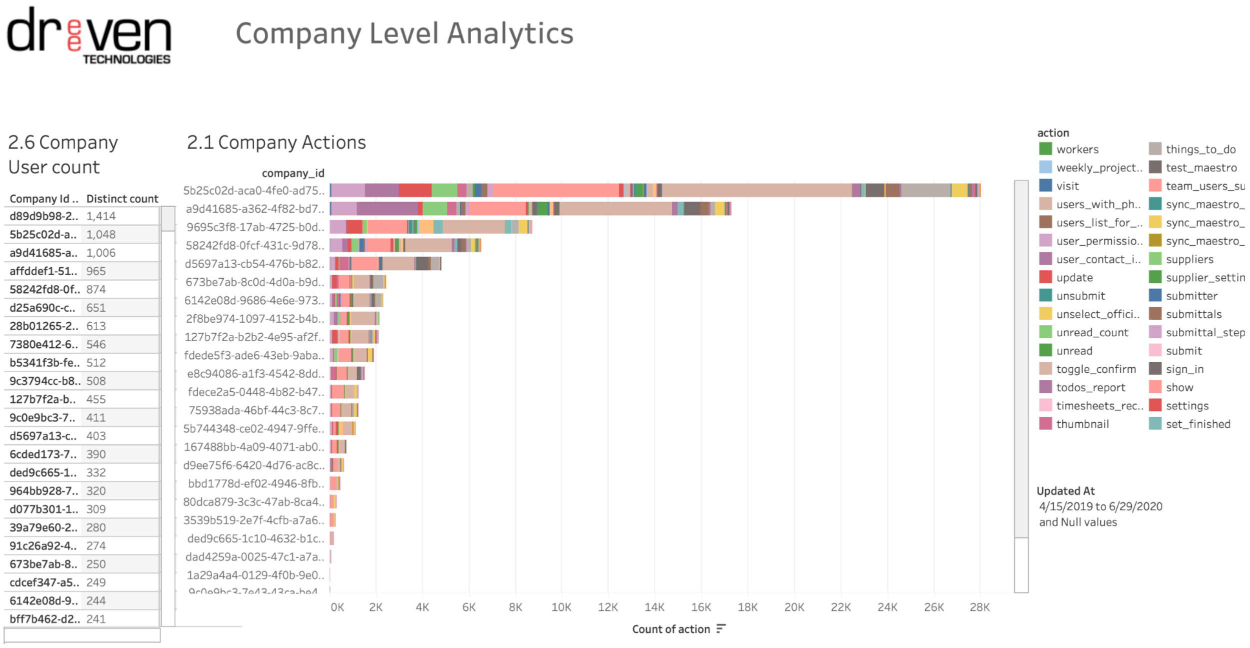
**Fig : Platform Level Analytics Dashboard**

In this dashboard, we are trying to use the data to analyse various aspects about the platform.

The top-left visualization shows us that the number of licenses that were created in 2023 were much greater than number of liceneses that were created in 2022. Similar pattern is observed for other years as well. This tells us that Dreeven as a platform is gaining popularity with time and that the company is doing something right as it is able to sell greater number of licences over time.

The donut viz shows the Dreeven team about the total number of users that they have on the platform. It also shows us the distribution of these users by the type of liceneses that they own.

Another visualization displays information about the type and quantity of license that each company holds.



Company Level Analytics

2.6 Company User count

| Company Id .. | Distinct count |
|---|---|
| d89d9b98-2.. | 1,414 |
| 5b25c02d-a.. | 1,048 |
| a9d41685-a.. | 1,006 |
| affddef1-51.. | 965 |
| 58242fd8-0f.. | 874 |
| d25a690c-c.. | 651 |
| 28b01265-2.. | 613 |
| 7380e412-6.. | 546 |
| b5341f3b-fe.. | 512 |
| 9c3794cc-b8.. | 508 |
| 127b7f2a-b.. | 455 |
| 9c0e9bc3-7.. | 411 |
| d5697a13-c.. | 403 |
| 6cded173-7.. | 390 |
| ded9c665-1.. | 332 |
| 964bb928-7.. | 320 |
| d077b301-1.. | 309 |
| 39a79e60-2.. | 280 |
| 91c26a92-4.. | 274 |
| 673be7ab-8.. | 250 |
| cdcef347-a5.. | 249 |
| 6142e08d-9.. | 244 |
| bff7b462-d2.. | 241 |

2.1 Company Actions

**Fig : Company Level Analytics Dashboard**

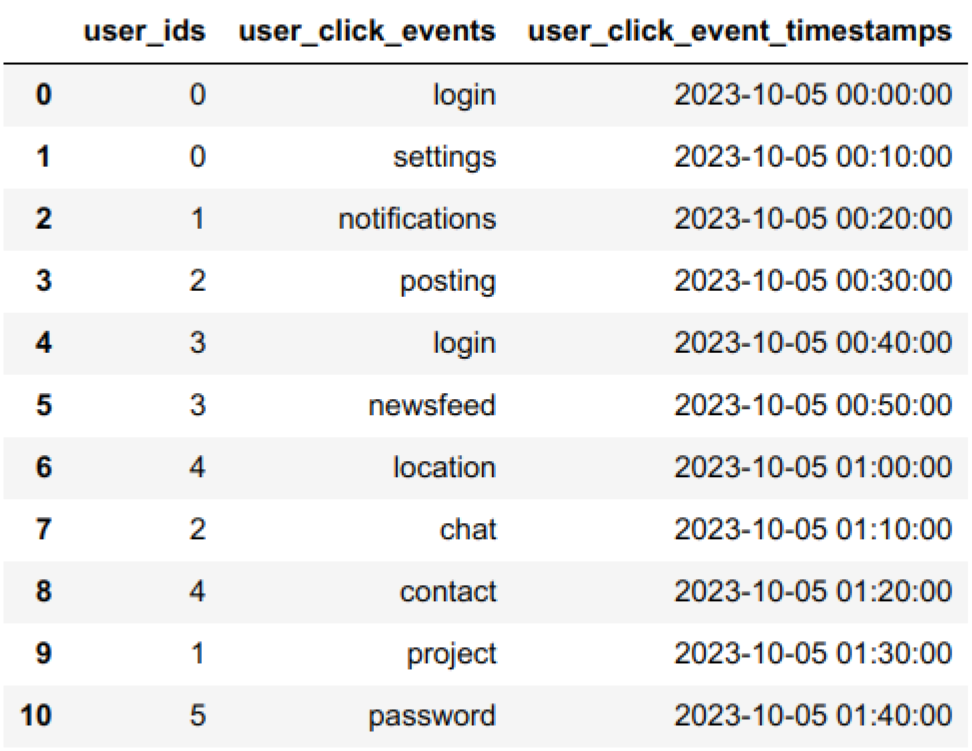
In this dashboard, the objective is to understand few statistics related to the companies that are doing business with Dreeven.

Viz 2.1 shows the features (actions) that are being used by any company and how often it is being used by them. It helps in finding out which are the features (actions) that any company finds to be most useful on the Dreeven Platform. This knowledge can be used to segment companies based on their usage on the platform.

Viz 2.6 shows the number of distinct users that are using the Dreeven Platform from each company. It tells us that more than one user can use the platform from any company and therefore, each of these users might behave in a different way on the platform.

# 5. Synthetic Dataset and Feature Extraction

To prototype the churn prediction framework, a synthetic dataset was initially employed for feasibility analysis. Feature extraction techniques were applied to capture key user interactions and temporal patterns within the clickstream data.

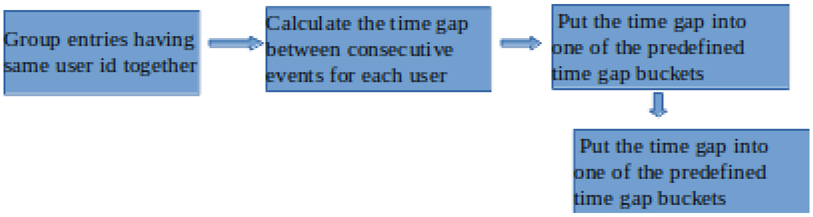


| | user_ids | user_click_events | user_click_event_timestamps |
|---|---|---|---|
| **0** | 0 | login | 2023-10-05 00:00:00 |
| **1** | 0 | settings | 2023-10-05 00:10:00 |
| **2** | 1 | notifications | 2023-10-05 00:20:00 |
| **3** | 2 | posting | 2023-10-05 00:30:00 |
| **4** | 3 | login | 2023-10-05 00:40:00 |
| **5** | 3 | newsfeed | 2023-10-05 00:50:00 |
| **6** | 4 | location | 2023-10-05 01:00:00 |
| **7** | 2 | chat | 2023-10-05 01:10:00 |
| **8** | 4 | contact | 2023-10-05 01:20:00 |
| **9** | 1 | project | 2023-10-05 01:30:00 |
| **10** | 5 | password | 2023-10-05 01:40:00 |

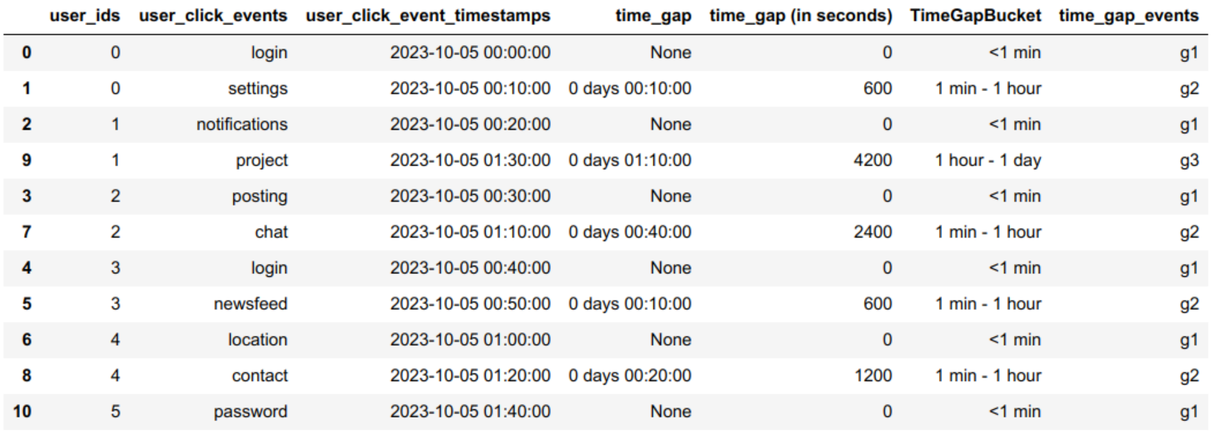
**Fig : Sample of the generated synthetic dataset**

# 6. Dataset Preparation Pipeline

The clickstream data was transformed into sequences of user events, with time gaps categorized into discrete intervals. This prepared the data for similarity analysis and subsequent clustering of user behavior patterns. The pipeline can be visualized using this block diagram :


Group entries having same user id together
Calculate the time gap between consecutive events for each user
Put the time gap into one of the predefined time gap buckets
Put the time gap into one of the predefined time gap buckets

The above synthetic dataset when passed through this data preparation pipeline, looks like this :



| | user_ids | user_click_events | user_click_event_timestamps | time_gap | time_gap (in seconds) | TimeGapBucket | time_gap_events |
|---|---|---|---|---|---|---|---|
| **0** | 0 | login | 2023-10-05 00:00:00 | None | 0 | <1 min | g1 |
| **1** | 0 | settings | 2023-10-05 00:10:00 | 0 days 00:10:00 | 600 | 1 min - 1 hour | g2 |
| **2** | 1 | notifications | 2023-10-05 00:20:00 | None | 0 | <1 min | g1 |
| **9** | 1 | project | 2023-10-05 01:30:00 | 0 days 01:10:00 | 4200 | 1 hour - 1 day | g3 |
| **3** | 2 | posting | 2023-10-05 00:30:00 | None | 0 | <1 min | g1 |
| **7** | 2 | chat | 2023-10-05 01:10:00 | 0 days 00:40:00 | 2400 | 1 min - 1 hour | g2 |
| **4** | 3 | login | 2023-10-05 00:40:00 | None | 0 | <1 min | g1 |
| **5** | 3 | newsfeed | 2023-10-05 00:50:00 | 0 days 00:10:00 | 600 | 1 min - 1 hour | g2 |
| **6** | 4 | location | 2023-10-05 01:00:00 | None | 0 | <1 min | g1 |
| **8** | 4 | contact | 2023-10-05 01:20:00 | 0 days 00:20:00 | 1200 | 1 min - 1 hour | g2 |
| **10** | 5 | password | 2023-10-05 01:40:00 | None | 0 | <1 min | g1 |

**Fig : Synthetic dataset post data preparation pipeline**

# 7. Feature Extraction and Clickstream Similarity Graph Creation

Feature extraction focused on identifying meaningful patterns (k-grams) within user adoption paths. Clickstream similarity graphs were constructed using advanced distance metrics to quantify the similarities between users based on their behavior sequences.

**Formatting the clickstream data into a sequence of events :**

In the prepared dataset, each user's clickstream activity is represented as a sequence comprising click events and time gap events.

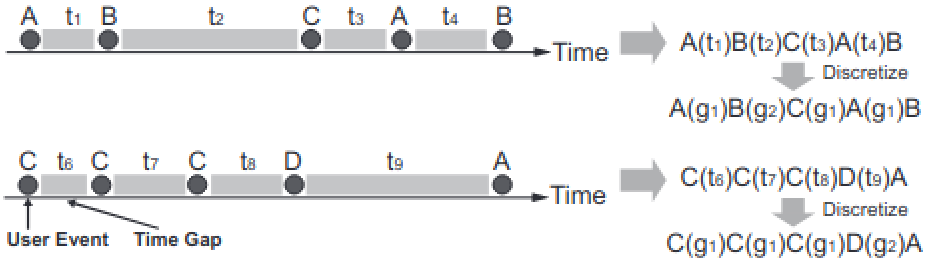
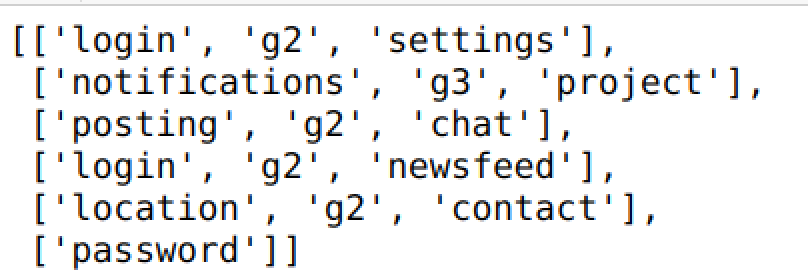

A
$t_1$
B
$t_2$
C
$t_3$
A
$t_4$
B
Time
$A(t_1)B(t_2)C(t_3)A(t_4)B$
Discretize
$A(g_1)B(g_2)C(g_1)A(g_1)B$
C
$t_6$
C
$t_7$
C
$t_8$
D
$t_9$
A
Time
$C(t_6)C(t_7)C(t_8)D(t_9)A$
Discretize
$C(g_1)C(g_1)C(g_1)D(g_2)A$
User Event
Time Gap

**Fig : User-activity data arranged as a sequence**

Above, we see how the activity of 2 different users are arranged to form a sequence. Here A,B,C,D represent click events where as g1 and g2 represent the time gap event. Below, we see the activity of all the users (user id : 0-5), expressed in the form of sequence of events.



```
[['login', 'g2', 'settings'],
 ['notifications', 'g3', 'project'],
 ['posting', 'g2', 'chat'],
 ['login', 'g2', 'newsfeed'],
 ['location', 'g2', 'contact'],
 ['password']]
```

To standardize our analytical approach, time gaps are categorized into five discrete intervals: <1s, [1s, 1min], (1min, 1h], (1h, 1day], >1day, denoted by 'g1', 'g2', 'g3', 'g4', and 'g5', respectively.

Visually, similarities between users such as 'u0' and 'u3' can be discerned.

**Feature extraction and Clickstream similarity graph creation :**

The methodology employed involves the extraction of subsequences from clickstreams to derive features for similarity comparison. Specifically, a clickstream is formalized as a sequence $S=(s_1,s_2,...,s_i,...,s_{n,})$, where $s_i$ represents the i-th element in the sequence, denoting either a click event or a time gap event, and n denotes the total number of events in the sequence.

The set $T_k$ is defined as the collection of all possible k-grams (k-consecutive elements) within sequence S

To compute the distance between two sequences S1 and S2 for a chosen k, the following steps are undertaken:

1. Determine the set $T=T_k(S_1)\cup T_k(S_2)$, comprising all potential k-grams from both sequences.
2. Calculate the normalized frequency of each k-gram within each sequence.

Subsequently, compute the normalized Polar Distance $D(S_1,S_2)$ between the two arrays:

$$\frac{1}{\pi}\cos^{-1}\frac{\sum_{j=1}^{n} c_{1j}\times c_{2j}}{\sqrt{\sum_{j=1}^{n}(c_{1j})^2}\times\sqrt{\sum_{j=1}^{n}(c_{2j})^2}}.$$

**Fig : Polar distance between $S_1$ and $S_2$**

The computed distance value, ranging from 0 to 1, indicates the similarity between the two clickstreams, with a smaller distance implying higher similarity. The Polar Distance metric is preferred over alternatives such as Euclidean distance due to its effectiveness in handling highly sparse vectors, emphasizing the "directionality" rather than the "magnitude" of vectors.

For the provided sample dataset containing six users (user IDs: 0-5) and utilizing a 1-gram feature representation for each user's clickstream, the resulting similarity graph is illustrated accordingly.

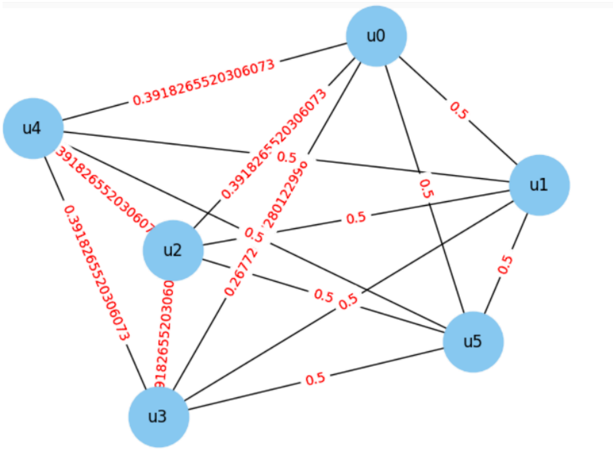


**Fig : Similarity graph created for all the users in the synthetic dataset**

## 8. Feature Pruning-Based Clickstream Clustering

To optimize user clustering, feature pruning techniques were applied using Chi-Square statistics to identify discriminative features. Divisive hierarchical clustering methods were then employed to group users into behaviorally similar clusters.

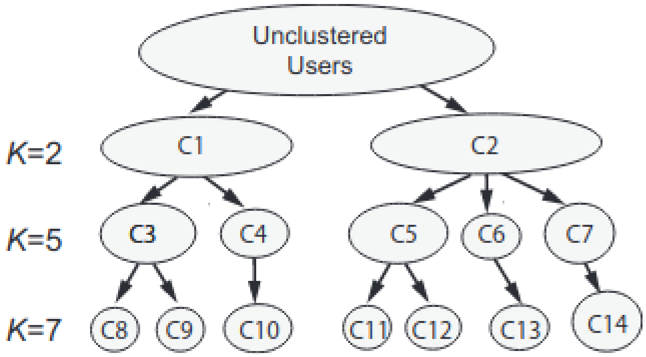


**Figure 1. Hierarchy of the behavioral clusters.**

# 9. Technical Challenges and Solutions

The internship project encountered several technical challenges, including scalability issues with real-time datasets and computational complexity during feature extraction. The real data can have much greater number of features than the synthetic data used. This could result in high computational cost during similarity graph creation. To solve this problem, the approach which was adopted for creating clusters was to use feature pruning technique. This can help us in identifying features that are most dominant (high Chi square value) and thus helps us select only those features which are useful for our application. This significantly brings down the overall computational cost.

# 10. Timeline and Project Phases

The internship project spanned six months and was divided into eight distinct phases:

1. Data acquisition and cleaning.
2. Literature review and methodology exploration.
3. Synthetic dataset implementation and validation.
4. Real dataset transformation and adaptation.
5. Algorithm development and feature engineering.
6. Implementation and testing of predictive models.
7. Integration with Dreeven Platform.
8. Knowledge transfer and documentation.

## 11. Results and Analysis

The data-driven framework demonstrated promising results in predicting user churn and identifying at-risk users. Analysis of clustering outcomes and model accuracy provided insights into user behavior patterns and actionable strategies for retention.

## 12. Conclusion and Reflection

In conclusion, the internship project achieved its objectives of developing a user adoption path data-driven framework for churn prediction. Reflections on the internship experience include personal growth, technical skills acquired, and recommendations for future research and development.

## 13. Recommendations for Future Work

Future work recommendations include:

- Continuous refinement and optimization of predictive models based on real-time user data.
- Integration of advanced machine learning techniques for enhanced churn prediction accuracy.
- Collaboration with domain experts and stakeholders to implement targeted user retention strategies based on behavioral insights.

# 14. References

1. Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2010). Factorizing personalized Markov chains for next-basket recommendation. In Proceedings of the 19th international conference on World wide web (pp. 811-820).
2. Benbouzid, B., Gao, X., & Wang, Y. (2016). Analyzing User Behavior in Clickstreams for Interaction Patterns. Retrieved from http://people.cs.uchicago.edu/~ravenben/publications/pdf/clickstream-chi16.pdf