FINAL PROJECT REPORT

Team D&L Labs

Luis Tun, Connor Davis, Dawn Elizabeth, Kevinray Lu

{tun, cjdavis, dawneliz, zeyun}@usc.edu

University of Southern California

Project Title	User-Specific Cryptographic Asset Recommendation System
Date Started	08-29-2022
Date Completed	12-05-2022
Project Sponsor	Dr. Anna Farzindar

Table of Contents

Executive Summary

As Non-Fungible Tokens (NFTs) have become prevalent innovations within the blockchain network and grown to be in demand, there is a major business opportunity and untapped sector for revenue. Recommendation systems have shown to impact revenue, drive engagement and consumption of content, and increase usage of platforms (Piyadigama & Poravi 2022). In order to drive legitimacy, liquidity and engagement onto the RMDS platform, we want to integrate an NFT recommendation system.

Our goal is to build a robust NFT recommendation system that correctly suggests and presents assets based on categorizations, search similarities, and overall popularities. We will be using natural language processing (NLP) and supervised machine learning techniques to cluster NFTs based on their respective database collections. Then, we will record the cluster and their corresponding features to be categorized in the model. Second, the implementation of search similarity computation using NLP will be performed by utilizing user recommendation data. Lastly, the synthesis of the recommendation system will be residual on obtaining NFT trending/popularity in respect to price filter categorizations from past/present data to bestow a set of cryptographic assets to users. We will integrate this design and recommendation system into a non-coding User Experience prototype.

D&L Labs is a start-up research group that provides NFT marketplaces with cutting-edge user-centric recommendation innovations to expand and empower user engagement, accessibility and revenue for set platforms.

Lean Six Sigma Project

This project utilizes the DMAIC methodology, which organizes the sequencing of activities in five categories: Define, Measure, Analyze, Implement, Control. Those are therefore the first five categories in this report. Given the nature of this project as an Applied Data Science project, this report contains a sixth category covering technical details on the machine learning approaches utilized as well as system implementation, and prototype & demonstration.

1. Define Phase

The define phase focuses on describing the problem and the underlying process quantifiably to determine how performance will be measured

1.1. Customer Research

NFT Questionnaire:

- **What do you know about NFTs?**

➔ Most individuals had a very limited understanding of NFT and associated to digital art and nothing more. They had not technical understanding on how they functioned, how to purchase, how to create their own or where to start as well.

- **What kind of NFT would you be interested in? Why?**

➔ Most felt comfortable only buying digital art forms of NFTs because that's what they are comfortable with or have knowledge about in this space. They feel a bit intimidated by NFTs, which prevents users from wanting to get involved in the NFT space.

- **What kind of NFT platform are you using?**

➔ Users primarily depend on mainstream competitors, such as Opensea and Rarible because of their influence and knowledge of across a big population of users.

- **Can you guide me through the NFT page, from the beginning to end? What did you enjoy or didn't? How can you experience improve?**

➔ Users wanted a direct and simple platform to use, where they could easily search for NFTs. Find similar ones to the NFT they like, whether it's based on price/type/creator

➔ Didn't enjoy it was hard to navigate and needed further assistance on navigating the

platform. They would preferred a help center, page, or AI chatbot

- **What would incentives you to get into NFT's? Why?**

➔ Monetary incentives seemed to be the biggest motivator for users
➔ Accessibility to create profits or small business on the platform

- **If using a recommendation system for NFTs, what features would you want the system to provide for a good overall experience?**

➔ User wanted to know that the creator was credible, so giving them some form of access to credibility or transparency would help improve security
➔ Wanted to have user story to know who the artist or creator is, this will help to form a virtual connection between buyer and seller

RMDS Stakeholder:

- **What are the issues you are facing in your current NFT system?**

➔ Individuals are unfamiliar with the RMDS platform, it's functionalities or usage.
➔ How we present NFTs and lack of recommendations for users

- **What do you want to improve on the platform?**

➔ Increase usage of NFTs and engagement from users on the platform

Through our stakeholder meeting/check-ins and customer research, we have discovered crucial pain points to design our system and develop a model to best address both customers' needs and stakeholder's problems. We will use this information to create a NFT recommendation and strategic user experience/user interface to drive engagement, build legitimacy, and create liquidity in the product.

1.2. Tools Application

The most valuable define phase tool was the project charter, as by guiding the creation of a concise, high-level view of all strategic elements involved in the project the charter allows for effective communication with stakeholders which in turns enables better feedback gathering. Such feedback gathering in the early stages, when the project is surrounded by uncertainty, is invaluable, as it helps prevent misalignment issues downstream. A clear presentation of responsibilities, limitations and milestones helps set the project on the right footing.

2. Measure Phase

The measure phase consists of using measures or metrics to understand performance and the improvement opportunity.

2.1. Process Mapping

To depict the workflow for this project various different process maps were designed. These process maps include a SIPOC (appendix 1), High Level Process Map (appendix 2), Common Process Map (appendix 3), Detailed Process Map (appendix 4) and a Functional Process Map (appendix 5). Generating those maps was fundamental in learning more about the inputs and deliverables for which each step in the workflow.

This was a crucial part in designing our workflow when integrating both models into a singular NFT recommendation system. The work completed on the SIPOC diagram served in making a task list, which was elaborated on the Detailed Process Map. This allowed the team to follow a clear flow for development and serve as “checkpoints” for various taks throughout the project.

2.2 The Vital Few

As NFT become increasingly in demand, we must gather information to help build our model to recommend NFTs to our users. We must find datasets that have NFT features to build our model and outputs. Additionally, we want to find data that can focus on the users and their interest similarity. We will find these two datasets to train our models and enhance our recommendation system.

We ultimately plan to recommend NFTs to users based on their interest (categories), pricing, and types. We will be using NFT description data to categorize the NFTs and user data to build a similarity score between users. We will use the user similarity as well to create a cold start when individuals do not have a user history on the platform.

2.3. Data Exploration and Preprocessing

For our framework, we required datasets that describe the NFT - the asset, pricing to be able to categorize the NFTs and a dataset that could find similarities between different users based on their education, ethnicity, health impairments etc.

For the NFT categorization, we could not find a lot of dataset, we ended up choosing a

cdataset that described the historical sales data of NFTs over the time period of 5 months.

```
In [5]: df.head()
```

Out[5]:

	Unnamed: 0	collection_slug	collection_name	collection_url	asset_id	asset_name	asset_description	asset_contract_date	
0	0	rarible	Rarible	https://opensea.io/collection/rarible	18214580	Daft Punk Never Die	Piece of art, Daft Punk, always in our hears.\...	2020-05-27T16:53:32.834583	https://ope
1	1	rarebit-bunnies	Rarebit Bunnies	https://opensea.io/collection/rarebit-bunnies	18276844	Rarebit #164 - Wax Off Bunny	🐰♥ When it comes to high kicks this Rarebit's...	2021-01-21T20:43:08.113711	https://ope
2	2	rarible	Rarible	https://opensea.io/collection/rarible	16911700	Meditation	Meditation by Diana.\n\nMinted only 20 NFT col...	2020-05-27T16:53:32.834583	https://ope
3	3	rarible	Rarible	https://opensea.io/collection/rarible	16986936	I´m OG	This is one of the first NFTs in human history	2020-05-27T16:53:32.834583	https://ope
4	4	chainguardians	ChainGuardians	https://opensea.io/collection/chainguardians	13382164	Celia B100 #105	One of the original androids created within th...	2019-11-17T21:00:18.404059	https://ope

We used NLP to pre-process all the textual data, changed all text to lowercase and removed non-alphanumeric characters, stopwords and contractions to keep only the useful information for classification.

Textual data before and after NLP processing :

```
In [7]: df['asset_description']
```

```
Out[7]: 0         Piece of art, Daft Punk, always in our hears.\...
        1         🐰♥ When it comes to high kicks this Rarebit's...
        2         Meditation by Diana.\n\nMinted only 20 NFT col...
        3            This is one of the first NFTs in human history
        4         One of the original androids created within th...
                                    ...
        108142               Subtle and beautiful pixellated glass.
        108143                        DeFi taking place on Earth
        108144    Hai. My name is Chicco Egoboo, and before you ...
        108145                                 League of Kingdoms
        108146    This is a digitalized version of one of the 10...
        Name: asset_description, Length: 108147, dtype: object
```

```
In [14]: df['asset_description']
```

```
Out[14]: 0                  Piece art Daft Punk always hears Size x
         1         When comes high kicks Rarebit back paws lethal...
         2                  Meditation Diana Minted NFT collectibles
         3                        This one first NFTs human history
         4         One original androids created within Chainguar...
                                       ...
         108142                    Subtle beautiful pixellated glass
         108143                            DeFi taking place Earth
         108144    Hai My name Chicco Egoboo ask short anything I...
         108145                                    League Kingdoms
         108146    This digitalized version one papercut artworks...
         Name: asset_description, Length: 108147, dtype: object
```

From this dataset, We used all the rows of data but removed irrelevant columns of data that wouldn't be useful for classification, we extracted 7 of the 16 features and processed the asset description using NLP tools

```
[ ] df2.head()
```

	collection_name	asset_id	asset_name	asset_description	event_quantity	event_payment_symbol	event_total_price
0	Rarible	18214580	Daft Punk Never Die	Piece art Daft Punk always hears Size x	1.0	ETH	0.070000
1	Rarebit Bunnies	18276844	Rarebit #164 - Wax Off Bunny	When comes high kicks Rarebit back paws lethal...	1.0	ETH	0.150000
2	Rarible	16911700	Meditation	Meditation Diana Minted NFT collectibles	1.0	ETH	0.001000
3	Rarible	16986936	I´m OG	This one first NFTs human history	1.0	ETH	0.000647
4	ChainGuardians	13382164	Celia B100 #105	One original androids created within Chainguar...	1.0	ETH	0.200000

For the Similarity use case, we initially started with the User Recommendation dataset which was a JSON dataset. We spent over 2 weeks working on the dataset, retrieving the chunks and converting them to csv format. We also worked with only 10% of the dataset but because the dataset consisted of mostly NaN values, we shifted to the StackOverflow dataset.

The dataset consisted of 83K+ rows of data and 47 features. We did not have to do any pre-processing but we were limited to 50k rows of data due to computational power.
We also extracted 14 features from the database that were more unique to a user and could be used to describe a user in detail.

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

Out[7]:

MainBranch	Employment	Country	EdLevel	DevType	OrgSize	Currency	CompTotal	CompFreq	Age	Gender	Ethnicity	Accessibility	MentalHea
I am a developer by profession	Independent contractor, freelancer, or self-em...	Slovakia	Secondary school (e.g. American high school, G...	Developer, mobile	20 to 99 employees	EUR European Euro	4800.0	Monthly	25-34 years old	Man	White or of European descent	None of the above	None of ab
I am a student who is learning to code	Student, full-time	Netherlands	Bachelor's degree (B.A., B.S., B.Eng., etc.)	NaN	NaN	NaN	NaN	NaN	18-24 years old	Man	White or of European descent	None of the above	None of ab
I am not primarily a developer, but I write co...	Student, full-time	Russian Federation	Bachelor's degree (B.A., B.S., B.Eng., etc.)	NaN	NaN	NaN	NaN	NaN	18-24 years old	Man	Prefer not to say	None of the above	None of ab
I am a developer by profession	Employed full-time	Austria	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Developer, front-end	100 to 499 employees	EUR European Euro	NaN	Monthly	35-44 years old	Man	White or of European descent	I am deaf / hard of hearing	N
I am a developer by profession	Independent contractor, freelancer, or self-em...	United Kingdom of Great Britain and Northern I...	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	Developer, desktop or enterprise applications;...	Just me - I am a freelancer, sole proprietor, ...	GBP\tPound sterling	NaN	NaN	25-34 years old	Man	White or of European descent	None of the above	N

2.4. Tools Application

Our tools can be broken down into two categories, algorithmic ones and project planning ones. For developing our models, we utilized data normalization tools, Principal Component Analysis, Singular-Value Decomposition, and data standardization tools from SK-Learn library. A detailed description of tools can be found in the "System Implementation" section. In short, those tools not only helped us regularize data into compatible forms/structures for model fitting in order to achieve a higher accuracy level, but also promoted framework performance in terms of time efficiency by reducing/reshaping dimensionalities for inputted features.

For project planning, process mapping (SIPOC, high-level, detailed, functional, etc.) played an important role for our group in particular. As we do have three major phases in our project (NFT grouping algorithm, user similarity computation & final framework combination, UI prototype design), a systematic planning for each stage of the project became essential to rely on for progress checking. We were able to efficiently follow the breakdown of inputs and outputs from steps we created and hence to finish our project on time thanks to this management tool.

3. Analysis Phase

The analysis phase focuses on identifying the true root causes of the underlying

problem.

3.1. Selecting Charts for Analysis

Aligned with the lean ideology, we followed a logical sequence of activities to identify root problems in the project. We selected two charts for root cause analysis. The first is the Fishbone Diagram helped in mapping which specific elements fed the problem prone areas. The second is the 5 Whys technique sheds light on the superficial problems and helps identify the true causes of error so that they can then be addressed.

3.2. Root Cause Analysis

In order to improve a system we must first know its current state. Our team used the well established Lean Six Sigma tools to depict the project's current state. We generated a Fishbone Diagram (appendix 6), which looks at the five sources of project defects: Dataset failures, Model Error, Model Assessment, Team Members, and Failure of Natural Language Processing. For each of these categories, the team brainstormed which elements could have potentially driven towards the effect (Y) being analyzed, which was "Failure to deliver the product (i.e. NFT Recommendation System)". The highlighted challenges from this exercise were (1) Data Collection and Preprocessing, (2) Complex Machine Learning Models, and (3) Model Evaluation (4) Team Member's skillset abilities.

Finally, we performed a 5 Whys (appendix 7) analysis, which deals with looking at the immediately visible problems that can cause the team to deliver "NFT Recommendations", and then asking why five times, in order to drive towards the root issue that is behind the superficially observable problems. Some common root causes were identified and are explored in further detail in the section below.

3.3. Sources of Variation

The analysis charts created evidence that the main sources of variation lie in four areas:

a. Data Collection & Preprocessing: Some of our problems occurred in the data collection and preprocessing for our user similarity score. We ran into issues where our platforms and devices could not read our original JSON file either because the file was too large or unable to merge when divided up to be machine readable. We had to pivot and find

other datasets to continue with our user similarity score. Another issue we had was find relative datasets since NFT are a new emerging field.

b. Machine Learning Models: Our project had two advanced machine learning models that needed to be combined, these models were the supervised machine learning and user similarity score. They didn't have any initial issues with merging but was difficult to find a similarity feature to combine both at first.

c. Model Evaluation: As the NFT is a relatively new field, we had difficulty establishing techniques that we wanted to implement in our recommendation. Other recommendation systems have varying research topics and are not focused on NFTs, which makes finding a base model difficult or what techniques to implement in our recommendation system.

d. Team Member's Skill Sets: We had a limited number of coders on the team, our team consisted of two coders and two designers. We tried to maximize each of our skill sets to provide the best possible product to our stakeholder. However, this was difficult as it put more pressure on trying to develop our models.

3.4. Potential Solutions

Some potential solutions to ensure proper data collection are:

- Setting appropriate time bounds to collect limited data
- Communicating with other teams, professor, or teaching assistance for potential support

Model Evaluation:

- Referencing baseline model to continue improving
- Researching further recommendation system models to build techniques

Team Member's Skill Set:

- Maximizing individual member's skill set from coding to design perspective
- Need to outsource or find a front end developer to connect the backend and UI design

3.5. Tools Application

The main tools leveraged in this stage was the root cause analysis. Employing this lean six sigma strategy provided invaluable guidance in identifying underlying issues and attending to those problems. Implementing this methodology assisted the team in building strong changes and overall performance of our model.

4. Improve Phase

Within the improvement phase the team identifies and assesses the validity of the model and tests the most adequate improvements that address root causes and increase performance of the supervised machine learning models and similarities.

4.1. Solution Evaluation

For each of the four critical areas for which divergent solution routes were generated, the team also evaluated their pros and cons, which are presented below.

Data Collection: The source of the data should be reliable, non-fraudulent and unbiased to make accurate decisions such as Kaggle datasets.

Methods	Pros	Cons
Kaggle NFT	This is a large dataset with various features to select from, can improve accuracy of model, and be diverse of model building.	Data needs heavy data cleaning and has too many features, can be hard to select and build an accurate model
Kaggle User Data	Large dataset that can improve accuracy	Older dataset and not specific to NFTs, can be limiting and not relevant

Data Preprocessing and Data Analysis: The data processing techniques we considered mostly revolved around increasing the contextual understanding of the features and user interest. We worked on understanding the importance of removing urls, emojis, stop words, contractions and other errors within the collective datasets. The machine learning model aims to recommend NFTs to users based on categories, popularity and similarities. Thus the need for concise analysis of features.

Data Preprocessing & Feature Selection Methods	Pros	Cons
Text-Preprocessing Methods (Removing Numbers, Special Characters, URLs)	This will help our NLP model to better understand the context of the text by removing the irrelevant characters. NFTs in the same collection share similar urls.	Various special characters like hashtags, urls can have dependent value which can be useful in providing accurate sentiments.
Text-Preprocessing Methods(Word Stemming and Lemmatization)	This will help our NLP model to better understand the context of the text by removing the irrelevant words.	In a few cases, this may lead to losing context and decrease in accuracy due to higher error rate.
Text-Preprocessing Methods (Removing stop words, contractions)	This will help our NLP model to better understand the context and semantics of the words by just keeping the base form.	In a few cases, this may lead to decrease in accuracy.
Feature Importance - TF-IDF	This automatically takes into account all the extracted features and marks them equally important and does not give any one feature a weighted advantage over the other.	This can lead to volatile feature sets being given equal importance and might lead to lower accuracy or ever changing results.

Missing Values Imputation	When executed properly, can lead to better knowledge extraction and conclusion	Missing values can cause loss of efficiency in the knowledge extraction process, strong bias if missing data or mishandled
Feature Selection - irrelevant and redundant information	Helps to remove the redundant and irrelevant features which may induce accidental correlations in learning AI used later	If not done correctly, can lead to bias and false correlation in learning AI
Imbalanced Learning - Undersampling	When resampling the data to balance mining AI, they are independent of the data mining AI applied afterwards	Imbalance datasets can lead to standard classification learning AI that are biased toward majority class & create higher misclassification rate for minority class instances

Model Selection: The model should be lite enough to run in real time and quick enough to provide predictions on time. The model should be robust and capable of capturing NFT recommendations based on user similarity and supervised machine learning. This can also be used to build a cold start system for users without any user history.

Machine Learning Models	Pros	Cons
(NFT Grouping) Supervised Machine Learning - price and currency as one-hot vector	Straightforward approach of just 3 vectorized features and 1 tag, easy to extract information from stakeholder's website.	Features do not share many visible similarities within the same collections. Hence results might not be optimal.

(NFT Grouping) Supervised Machine Learning - Collection tags as label and asset url added to the feature set	Potentially increases the accuracy when fitting, since the NFTs belonging to the same collection would have similar URL.	The format of urls for NFTs of the same collection are different on stakeholder's website. This approach would be limited to OpenSea unless RMDS chooses to have similar urls in the future.
(NFT Grouping) Unsupervised Machine Learning - Using all extracted features even collection tags	Avoids having too many clusters. Maybe the most reliable since NFTs of the same collection may not necessarily share too many common features.	Harder to validate, the overall training process would be longer. The NFTs can also not be recommended as part of a collection.
(User Similarity) **Cosine Similarity**	Has higher positive correlation and easily identifies how similar two documents are in terms of subject matter.	Not optimal when there are very less features and densely populated
(User Similarity) **Euclidean Distance**	Effective when collinearity exhibited.	Loses sensitivity when more features are added and are sparsely populated
(User Similarity) **TF-IDF Vectorizer**	Focuses on both frequency and importance of a word	Fails to provide linguistic information about words
(User Similarity) **CountVectorizer**	Finds statistically significant word	Inability to identify importance of a word, relationships between words

Model Evaluation:

Methods	Pros	Cons
Accuracy Score / F1 Score	Straightforward. The most prevalent method for checking performance of supervised Machine Learning algorithms.	Cannot be applied to unsupervised machine learning algorithms. Hard to compare to existing research projects using different datasets for sake of baseline models.
Silhouette Score	Reliable for unsupervised learning. Presents detailed progress in terms of MSE for different K's in K-Means Clustering.	MSE increases with increased training samples inevitably. Need to retain necessary features and minimize the score simultaneously.
Elbow Method	Reliable for unsupervised learning. Provides clear representation of a "best" K number in K-Means clustering where the "elbow" resides.	Not realistic to compare with existing works. Validation metrics need to refer to solid examples of K's implemented in past experiments.

4.2. Recommended Solution

Data Collection and Preprocessing:

Action Items	**People Responsible**	**Deadlines**
Data Collection & Project Design	Luis, Connor, Dawn, Kevinray	09/09/22
Data Cleaning / Preprocessing	Dawn	09/21/22
Feature Selection	Luis, Connor, Dawn, Kevinray	09/27/22

Machine Learning Model Development:

Action Items	**People Responsible**	**Deadlines**
Building Framework for Categorization	Kevinray	10/11/22
User Similarity Test	Dawn	11/01/22
NFT Recommendation System	Kevinray	11/08/22
Price Filtering & NFT Recommendation	Kevinray	11/21/22

Model Evaluation and UI design:

Action Items	**People Responsible**	**Deadlines**
UX Brainstorming	Connor & Luis	10/25/22
UX Sketch & Variables	Connor & Luis	10/28/22
UX/UI Beta Prototype	Connor & Luis	11/08/22
UX/UI Alpha Prototype	Connor & Luis	11/15/22
Project Closeout & Final Presentation	Luis, Dawn, Connor, Kevinray	11/22/22

4.3. Pilot Design

Beginning with the data exploration phase to the machine learning model building phase, testing and evaluation to the UI structuring has been the pilot phase of the project. After clear assessment of the Voice of Customer phase and feedback/expectations of the stakeholder the group began concise development. These assessments aided the group in staying on track to enable success in all stages of the project. After building the model and UI prototype throughout our work cycle we stayed in communication with the stakeholder alongside new customer feedback. In analysis of the models built we have combined both supervised machine learning models and similarity scores and tested their implementation. This implementation includes dividing users into two groups, gathering data from click history to get URLs of NFTs users visited, acquiring needed features from URLs, then fitting features to predict preferred collections.

Transitioning to the UI design, the UI reflects these predicted collections within the homepage. There is a 'Trending' page which reflects a cold start of collections in which are most popular, and a 'For You' page which reflects collections recommended for a new specific user. For user 24 within the UI, the team has also incorporated a price filter that displays NFTs within a fixed price range.

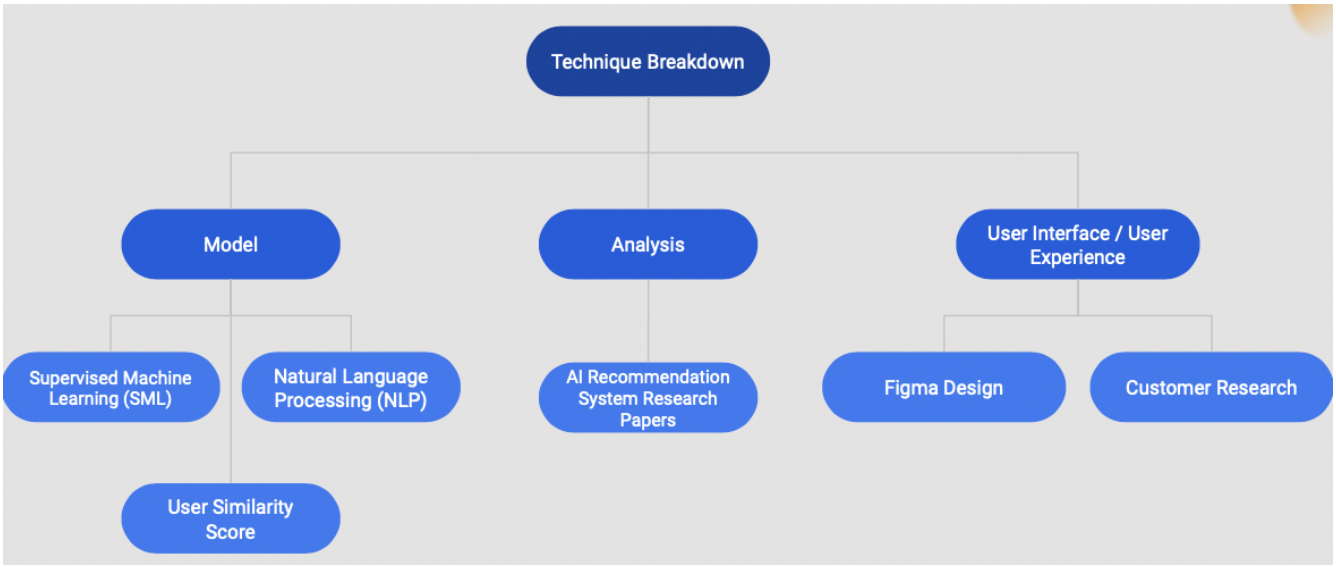
4.4. Work Breakdown Structure

The tasks involved in improving the project's deliverable can be divided into three main sections:

1. Model: We divided our models into three parts, which include a supervised machine learning (SML), natural language processing (NLP), and the user similarity score. We combined the supervised machine learning and user similarity model to create a NFT recommendation system. We used the NLP model to process our datasets.

2. Analysis: In this section, we researched various recommendation systems to create a baseline model and find techniques to implement in our own recommendation model.

3. User Interface / User Experience: We used design tools, such as Figma, and customer research to design an interactive non-coding prototype for the stakeholder.



5. Control Phase

The control phase handles the creation of sustainment strategies that ensure process performance maintains the improved state.

5.1. Control Solutions Considered

For setting up guardrails that continue the process improvements, we considered the below two solutions:

1. **Mistake Proofing (Poka-Yoke):** Since each touch-point is an opportunity for errors to be introduced, we focused on removing most of the work from the end user.
2. **Issue Tracking:** For tracking of any bugs or errors introduced in the system we have used open source platforms. These platforms should also ensure that package updates should not malfunction the system.

5.2. Control Solution Implementation

To maintain continuity in the deliverable's high performance, below control solutions have been implemented:

1. **Issue Tracking:** For achieving issue tracking we have used an agile progress log that provides a feedback channel for reporting and logging bugs / known issues and to focus on important tasks and keep plans up to date simultaneously.
2. **Documentation:** Maintaining and updating documentations provides a useful guide for management in how the new processes work. Documentation also includes preservation of project charters, process maps, customer needs and requirements, and charts and graphs created for the project.
3. **Mistake Proofing (Poka-Yoke):** Mistake proofing was implemented at two-levels. At UI level, we conduct customer research and coder feedback review to implement changes and accurate outputs. At the data level a data cleaning pipeline has been cleaned and revised to ensuring reproducibility.

6. Result and System Implementation

6.1. Machine Learning Approaches

Below shows a list of procedures taken/outputs obtained for each of the Machine Learning implementations we did using code. Please see reference page for all quoted/paraphrased citations.

1. Supervised / Unsupervised Machine Learning Model (NFT Grouping):

- **Vectorization**

To fit features into our framework, we need to vectorize chosen textual features such as asset description and payment method. For the payment method field in particular, we chose one-hot encoding to represent categorical data. For asset description, we encode the texts using tensorflow_hub.

- **Data Normalization**

Due to different encoding methodologies introduced above, we perform normalization across different features for them to have the same variance. Normalization will allow us to conduct higher learning rates and be less careful about initialization. It acts as a regularizer and in some cases, it eliminates the need for dropout and achieves accuracy with 14 times fewer steps and beats the original model by a significant margin (Ioffe et Szegedy, 2015).

- **Data Standardization**

Since we plan on using K-means clustering as part of the model, standardization needs to occur because it may strongly affect the performance of k-means (Su et al, 2009). This is particularly important when our datasets have different scales.

- **Principal Component Analysis**

Principal Component Analysis (PCA), or regularization of the data, acts directly between features and labels so it is more useful for models that describe the label based on the given features. As for PCA, it only considers the variable between the features (Kozakm 2018). We plan to use PCA for our model as we are only interested in using it as a compression technique for our high-dimensional feature set. Compressing the dimensionality of data would provide an even distribution of weight across different features, since they initially were of various lengths due to encoding.

- **Reshaping**

We do not reshape our data in-place. This step is to fit multiple high-dimensional data to the framework, alongside the machine learning frameworks we chose to only take the data input of two dimensions. Utilizing PCA, we were able to transform original features into vectors of the same lengths. Thus, reshaping is to flatten lists of features for sake of fitting without losing influence from features with shorter lengths.

From our implementation of above procedures, our team achieved above 97% accuracy level using SVM framework, exceeding our baseline model (Song et al, 2018). For the unsupervised model using K-Means, we also relied on past researches for numbers of K to choose from (Pham & Nguyen, 2005).

2. User Similarity:

- **Similarity Matrix**

For our implementation we compared the cosine similarity matrix and Euclidean distance to find the user similarity based on the basic user information. Cosine similarity gave better results in comparison to Euclidean distance as it had higher positive correlation and found the characteristic similarity to our

dataset and identities with higher accuracy.

- **Vectorizer**

For the language aware comparison, we compared the TF IDF and count vectorizer with the similarity matrix to see which gave us better results. CountVectorizer only focused on the frequency of the word and not the contextual importance, We hence chose TF IDF as it focuses on both the frequency and the importance of the text.

- **SVD Decomposition**

Given the limitations in terms of processing capabilities, we used Singular Value Decomposition to perform dimension reduction and at the same time to also not lose the useful properties of the matrix formed allowing us to express our original matrix in terms of linear combination of lower ranking matrices.

6.2. Final Model Implementation:

- **Model combination**
  Built a framework to combine User ID from the NFT cluster and User Similarity Scores (masking existing user IDs to clicking history dataset generated by RMDS in previous research). For experienced users, select the three most frequently predicted collections to recommend. For new users, select the three most frequent subsets of predictions from similar users.

- **Default settings**
  For experienced users, find features of NFTs according to users' clicking history URLs, and fit features into the NFT Grouping Framework to find collections of products to recommend.

- **Cold Start integration**
  For new users, find candidates that are similar (by user similarity score) to the current user while having non-empty clicking history entries (experienced users only). Then recommend the current user's corresponding collections resulting from other users.

- **Add-ons**
  The backend price filter can be realized with two inputs of lower range/higher range. Implemented trending recommendations using Python based on the number of transactions generated by various collections.

6.3. Prototype & Demonstration

The system functionalities on every page is detailed below:

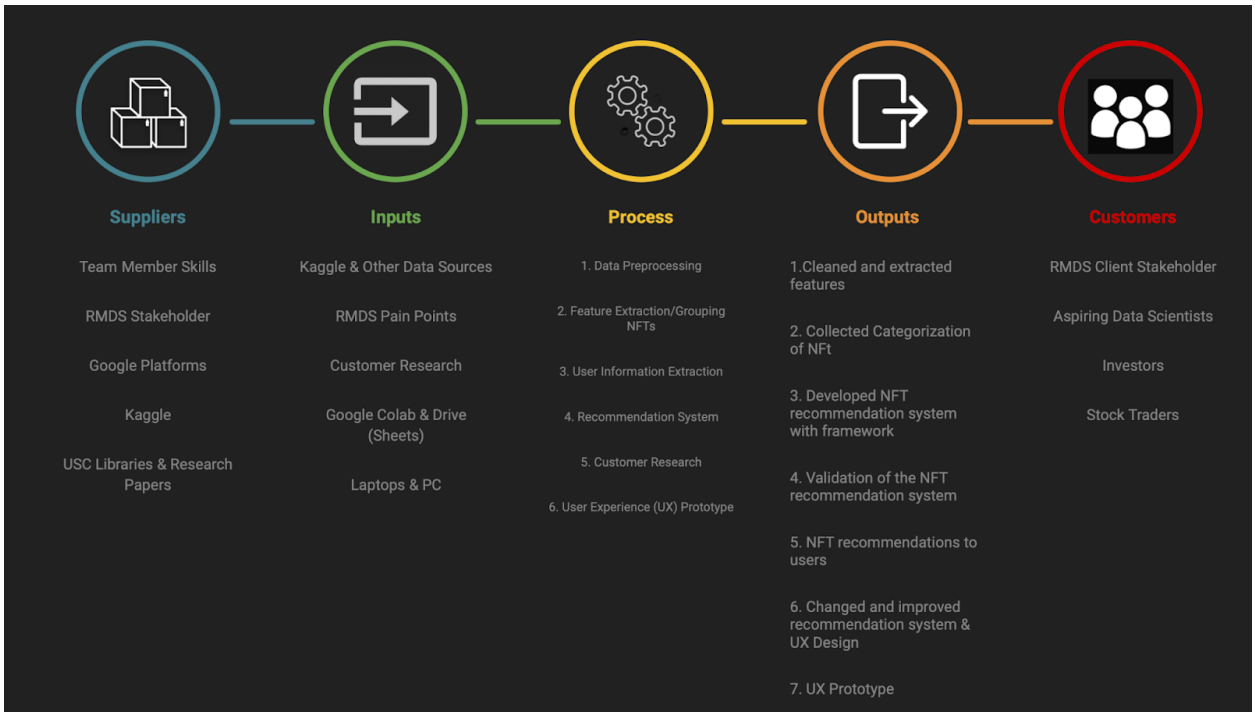
1. User 2052 clicks homepage to login; after logging in, three code emulated collections titled

'Sorare', 'Rarible', and 'Hashmasks' are shown under 'Trending NFTs'. The user then proceeds to click on a trading card from the Sorare collection titled 'Nikola Maksimovic'.

2. User 2052 then clicks on the 'For You' tab to display personalized NFTs from the machine learning methods executed; after that the profile tab is clicked to view their personal ID number then recedes back to the login page.
3. User 24 clicks homepage to login; after logging in, three new collections generate with price filtering once again emulating coding methodology. User 24 clicks another NFT from the Sorare collection before proceeding to the For You tab once again.
4. User 24 clicks on a NFT from the Rarible collection taking them to a user profile where that said NFT can be purchased; user 24 adds the NFT to cart alongside messaging the user for more info.
5. Figma Link: https://www.figma.com/proto/Wu84BqvYdDsICF8jSEv0IX/DSCI-560-Project?node-id=32%3A4&starting-point-node-id=32%3A4&scaling=scale-down

Appendix

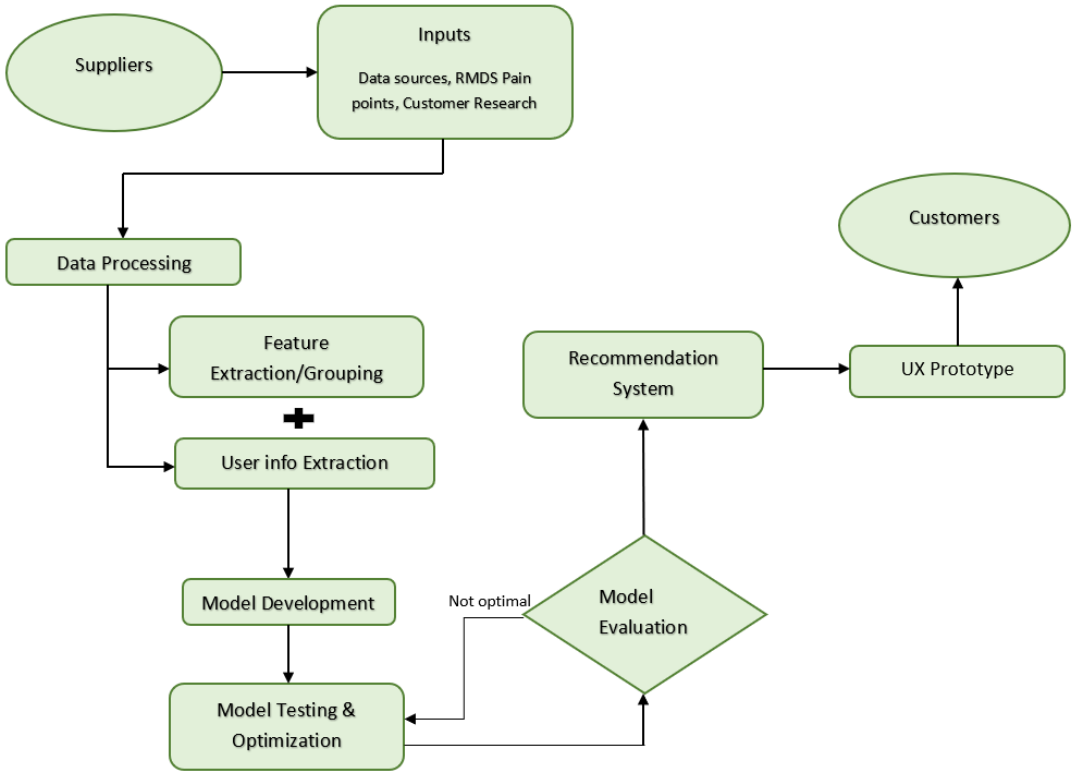
Appendix 1: SIPOC



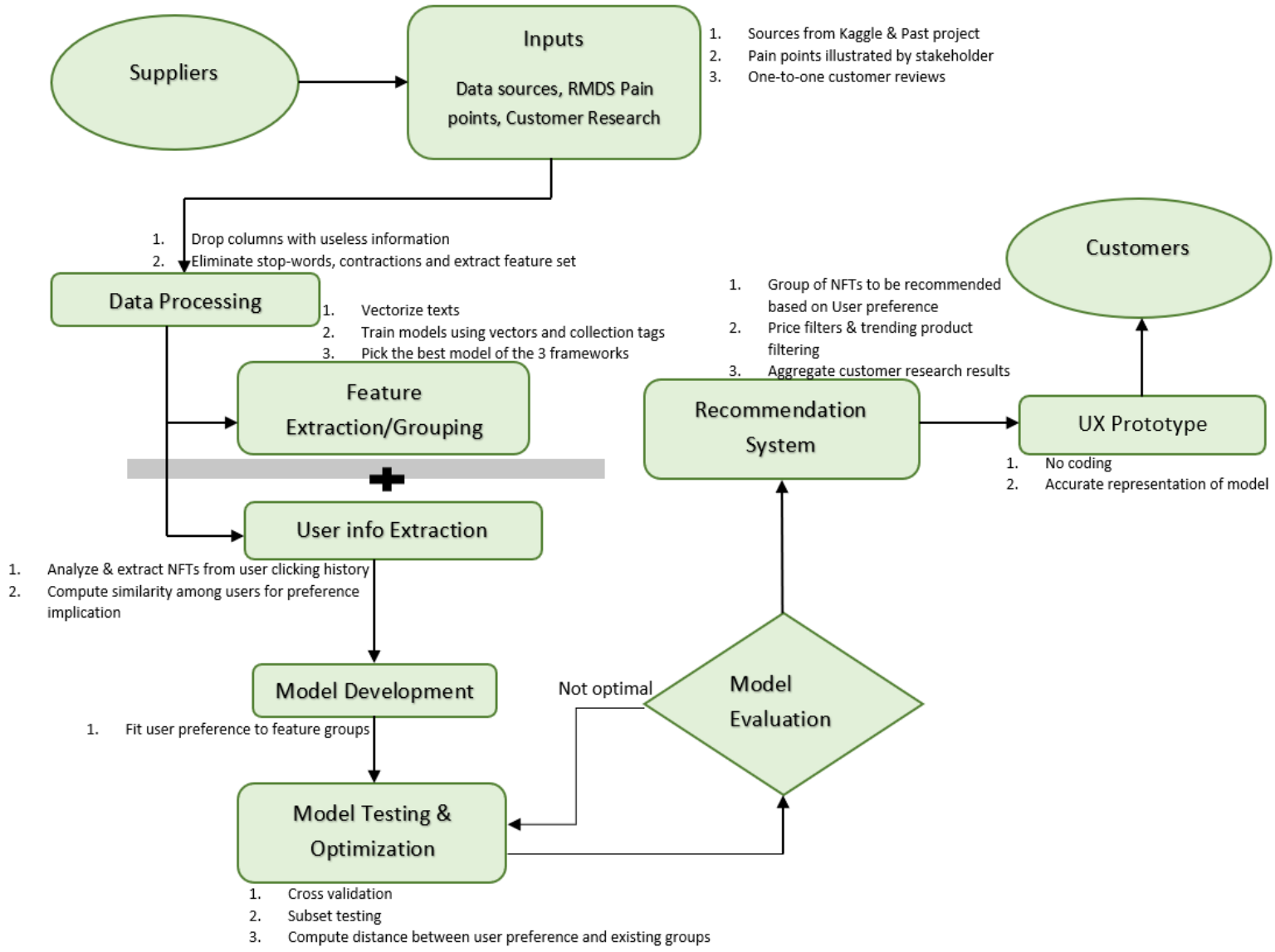
Appendix 2: High Level Process Map

Suppliers	Input	Process	Outputs	Customers
Team Member Skills RMDS Stakeholder Google Platforms Kaggle USC Libraries & Research Papers	Kaggle & Other Data Sources RDMS Pain Points Customer Research Google Colab & Platforms	1.Data Preprocessing 2.Feature/Categorizing NFTs 3.User Informations Extraction 4.Model Development / Testing 5.Recommendation System 6. Customer Research 7. UX Design	1.Cleaned and extracted features 2. Collected Categorization of NFT 3. Developed NFT recommendation system with framework 4. Validated NFT recommendation system 5. NFT recommended to users 6. Changed & improved to recommendation system & UX design 7. UX Prototype	RMDS Client & Stakeholders Aspiring Data Scientist & NFT fans Investors Stock Traders

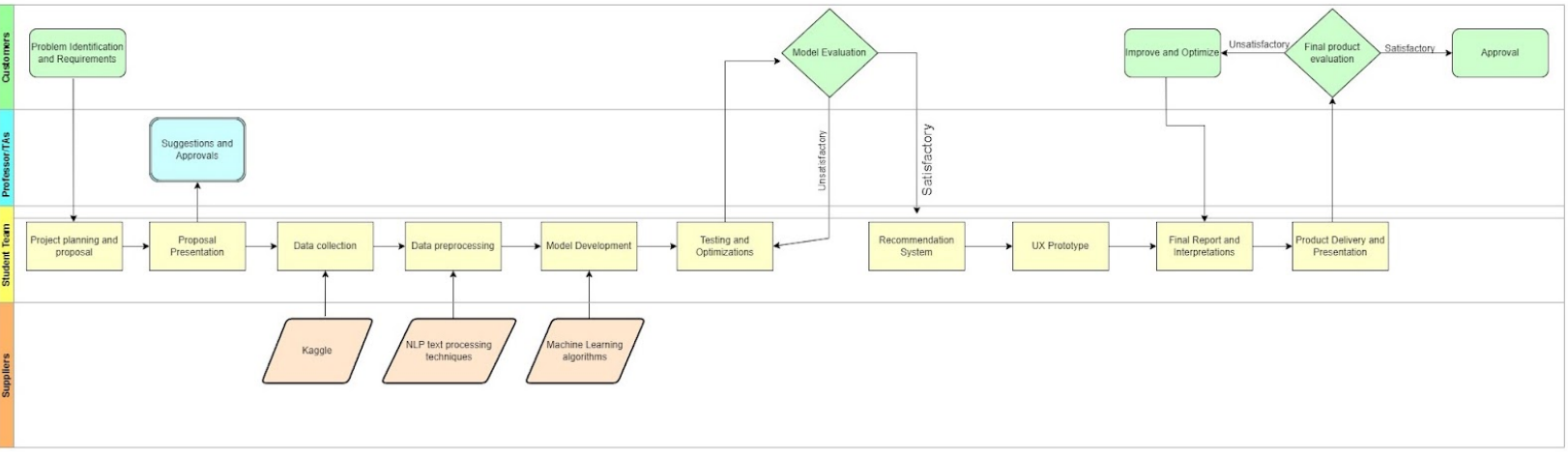
Appendix 3: Common Process Map



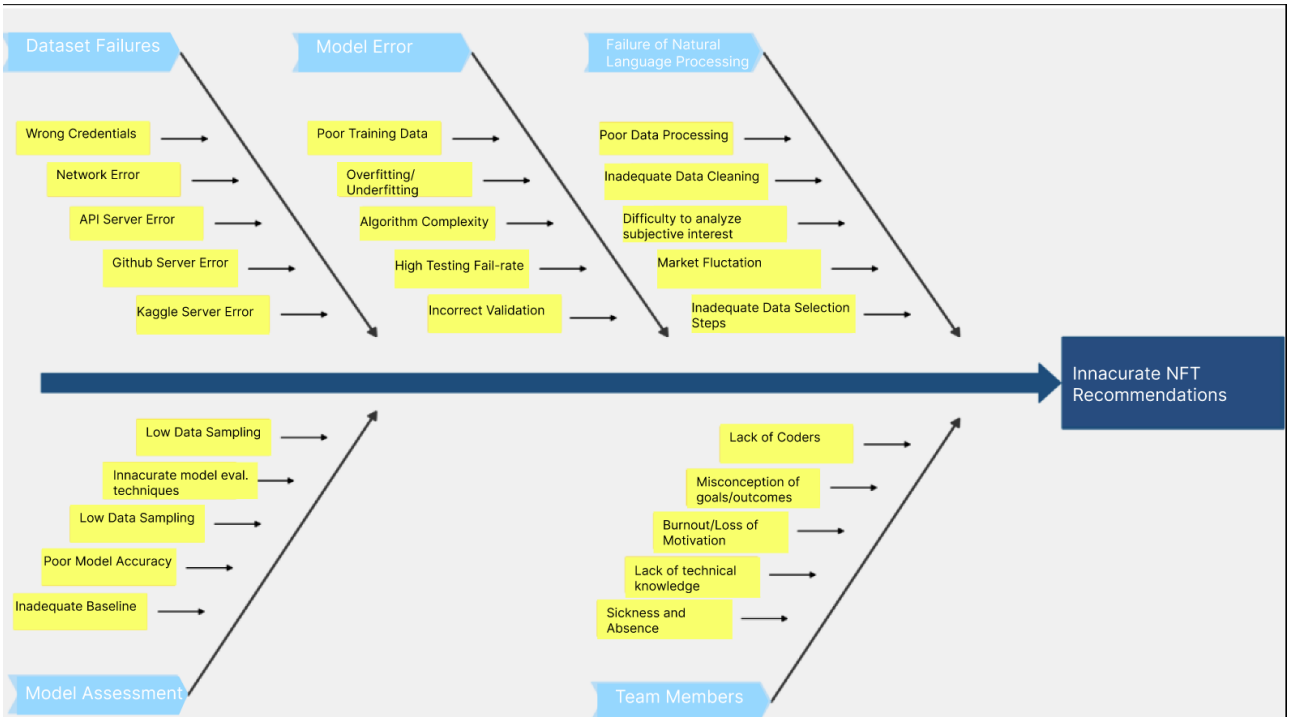
Appendix 4: Detailed Process Map



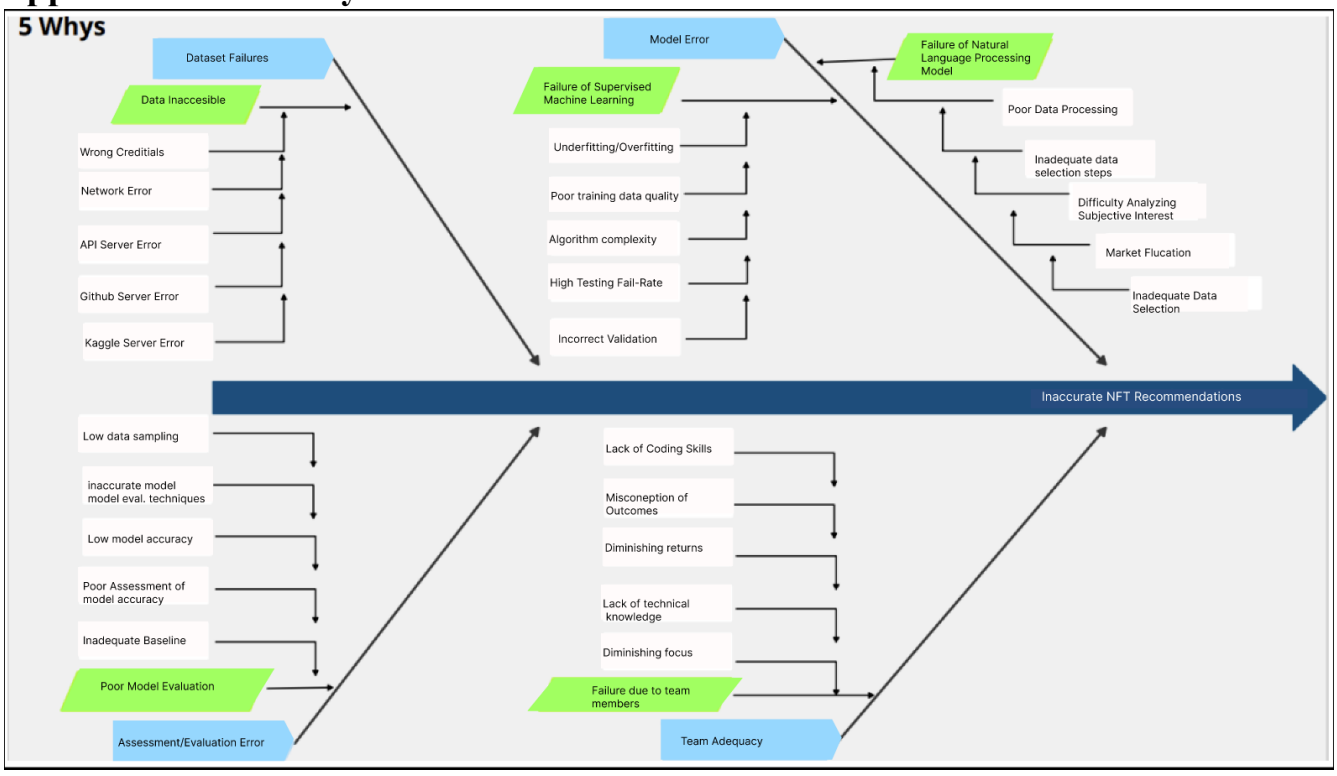
Appendix 5: Functional Process Map



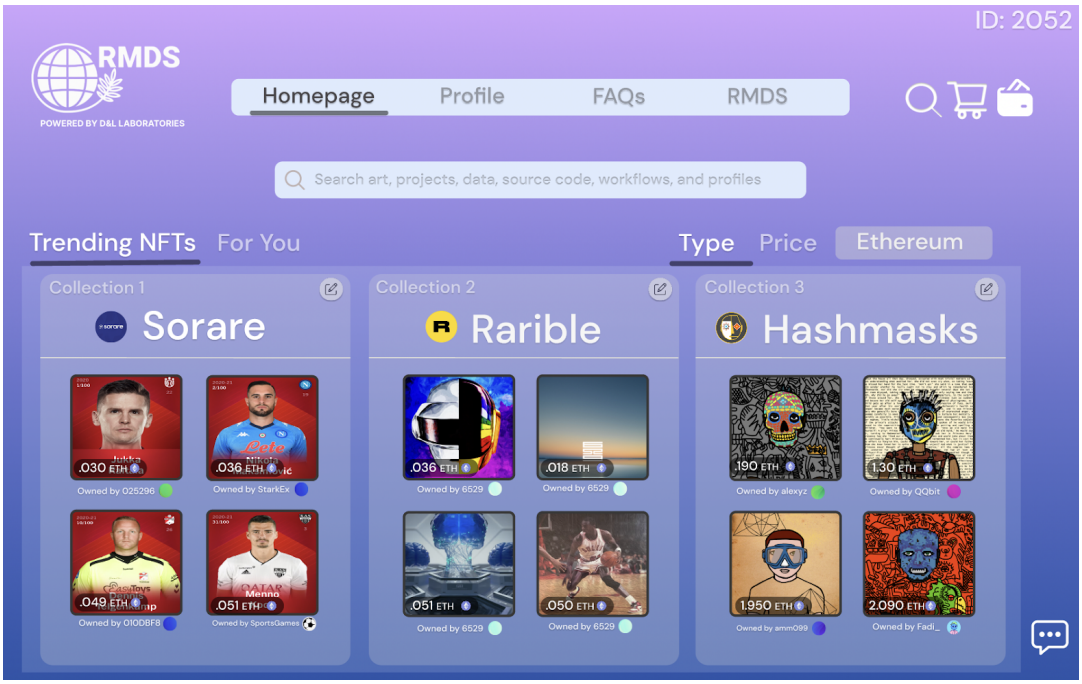
Appendix 6: Fishbone Diagram



Appendix 7: Five Whys

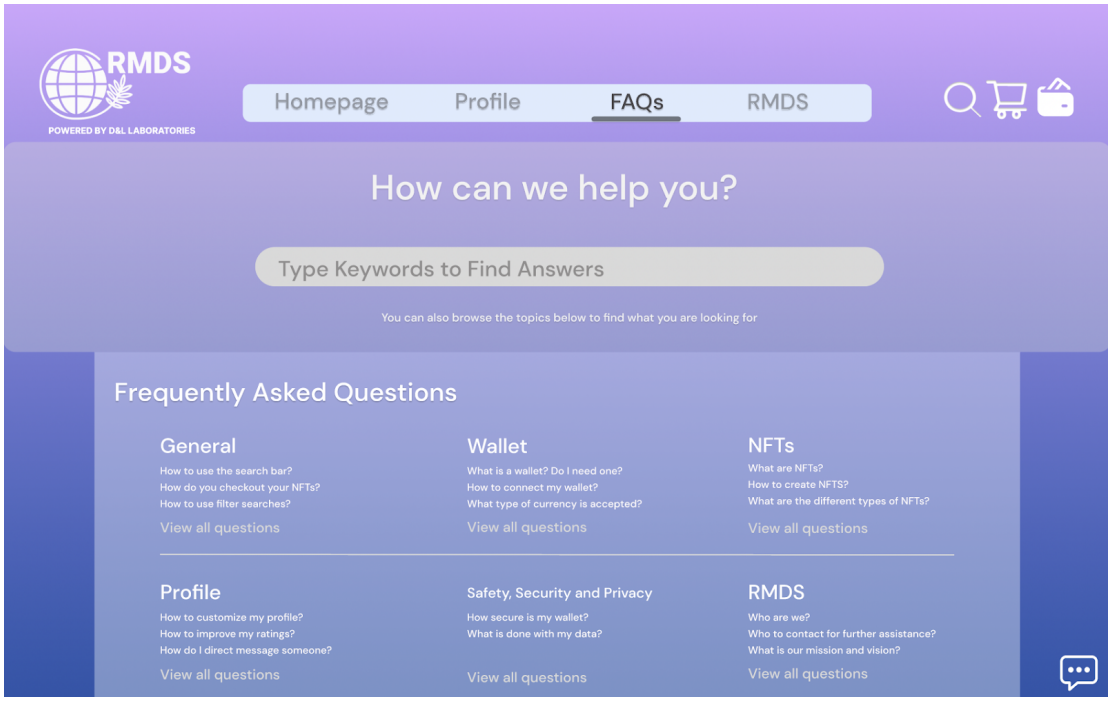


Appendix 8: Home Page



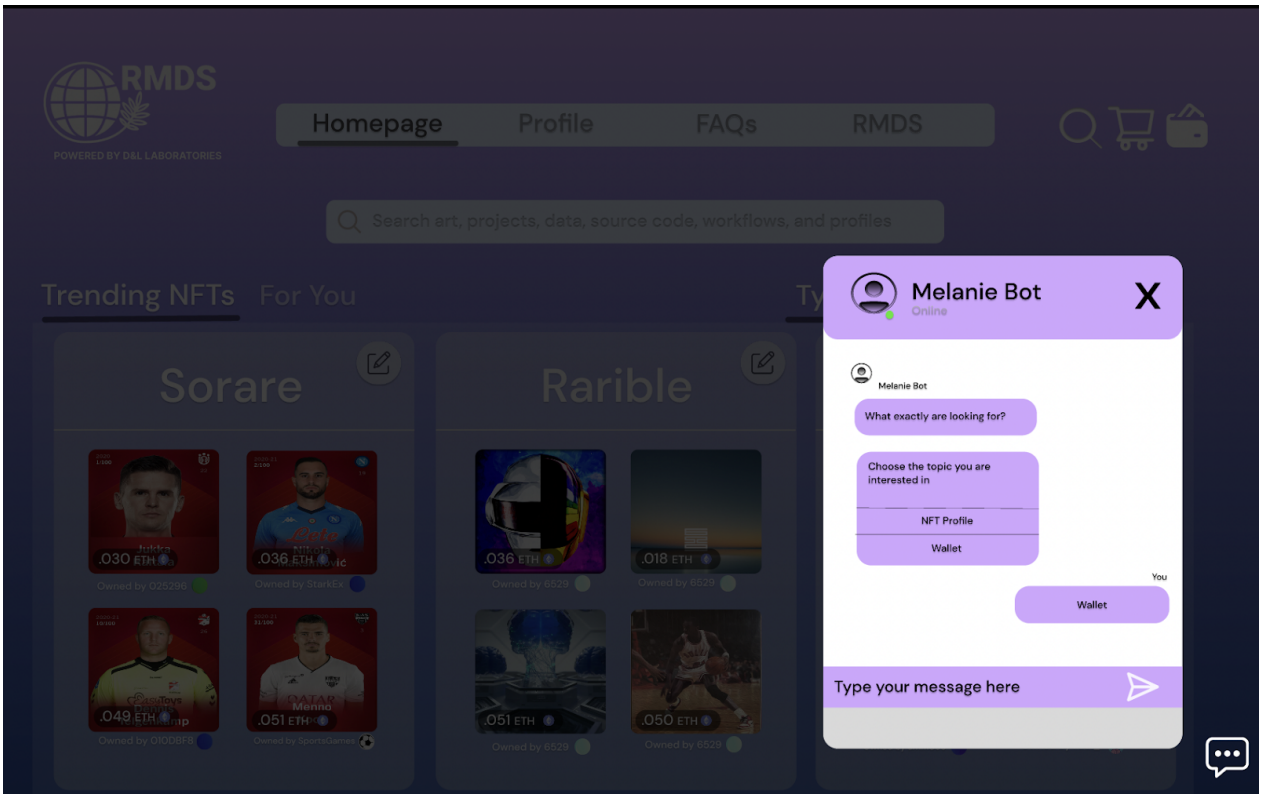
Appendix 9: User Profile

Appendix 10: FAQ Page



Appendix 11: RMDS Page

Appendix 12: Chatbot



Appendix 13: Wallet