NFT Price Prediction

Final Report
Usable AI Spring 2023

Abhijit Nayak
Andrea Chung

May 2, 2023

Data Collection

There was no change to our data collection method as same four csv files were used. As mentioned in the midterm report, none of the features were collected manually. The datasets are available on https://bitgrit.net/competition/17 and can be downloaded after the registration. The dataset is constrained to only personal and academic uses.

Data Management

To make improvements from prior data management, we retained target variable, "last sale price", separately to avoid the problem of having missing values after joining and merging process of datasets. We concatenated both the given train and test data and merged collections.csv and collections_twitter_stats.csv files by collections_id on left join.

Then we performed feature engineering to create more feature variables. Creation year, month and last sale year, month features were extracted from creation_date and last_sale_date column. Also, difference in days were computed as another feature from last_sale_date and creation_date. We encoded boolean type variables as numerical variables as well as encoded the categorical variables such as contract_type and verification_status into numerical variables too.

Furthermore, statistical feature engineering was implemented on the columns: rarity score, total supply, number of traits, and seller fees by calculating mean, median, maximum, minimum, sum, and standard deviation as individual features for each of the 4 columns. All the statistically engineered features were merged into the main dataframe. Then, the main dataframe was segregated into train and test sets based on their original lengths.

Finally, the target variable, "last sale price", was highly skewed, so we performed log transformation to transform it into approximately a normal distribution or to reduce skewness.

Analysis

This project's primary goal is to forecast NFT pricing. In the crypto market, it is becoming much more popular. Naturally, consumers would prefer to know the price of NFTs in advance so they may plan their transactions for NFTs to maximize profit. We have a variety of criteria, starting with social media-based features and group traits like rarity, seller fees, etc., to incorporate into the prediction model.

As a result, employing these features in a certain method (combining all the features into a single dataset) will aid the predictive models in learning the most important and useful patterns from the data needed to predict the pricing of NFTs. Since the problem description relates to a predictive task, forecasting or a straightforward regression technique could be used to solve it.

After observing the results from previous experiments mentioned in midterm report, there was huge scope to revamp the data pre-processing steps, add new features, remove bias and perform the machine learning modelling in a better way. So, we improved our models by using K Fold cross-validation approach to prevent overfitting with K as 16. This process splits the dataset into 16 folds. Here, the model was validated on the Kth fold and was trained on the remaining K-1 folds. The performance metric used to evaluate the models was average RMSE across all the folds. Lesser the RMSE, better the model's performance is.

In total, 3 regressor models--CatBoost (RMSE: 8.77), XGBoost (RMSE: 9.24), and LightGBM (RMSE: 10.18) performed better after experimenting with multiple models. As mentioned earlier, RMSE was used as the evaluation metric. We tried randomized search, grid search and bayesian optimization techniques for hyper-parameter tuning and observed that Bayesian Optimization gave better results because of the data-centric approach it follows.

During the model evaluation phase, we compared the average RMSE scores of all the 3 models mentioned above and selected CatBoost Regressor as the best model because it had the lowest average RMSE. Comparing to the prior experiments (midterm report), the choice of the prediction model changed from XGBoost to CatBoost with all the improved steps incorporated in this experiment mitigating the issues and errors.

We did not perform any data subsetting or subgroup analysis as our dataset was of medium sized and our systems were able to handle the complete load during the analysis phase. We didn't perform any statistical tests as we implemented tree based models only not the statistical models like Linear, lasso regression. Apart from this, the number of tests/experiments we performed were 2 experiments in total: first being explained in the midterm report and the second one was the improved stage being discussed in this final report. These experiments cover all the steps starting from data pre-processing, feature engineering to modelling and tuning.

One observation we had was that there was very small difference in the average RMSE scores of CatBoostRegressor (8.77) and XGBoostRegressor (9.24). So, there's scope to improve

the XGBoost regressor's scores by improving on the existing hyper-parameter tuning process of XGBoost.

Results

Models	RMSE
CatBoost Regressor	8.778
XGBoost Regressor	9.240
LightGBM Regressor	10.187
Bagging Regressor with Base estimator as LightGBM	12.47

Table 1. Result of Model Performance

Looking at the final RMSE scores for each model (Table 1), CatBoost regression model had the lowest RMSE value. Even after hyperparameter tuning, the CatBoost model had the best performance, which we selected at the end. After the final evaluation and selection, we performed model explanability and feature importances. The feature importances can help us to view the significant feature variables that effects the increase in model performance. The list of variables that plays important role in predicting the price of NFTs are obtained from feature importance analysis. The further explanation will be given in the argument section of the report.

Model Explanability

Explainable AI plays a significant role in the analysis and evaluation phase of the models. It gives a clear picture of how the model is making decisions and provide us with transparency of the model, which ensures not making any biased decisions to reach the predictions at final stage. It also helps to build trust in the AI models being developed. The following few line of code generates an interactive explainable html dashboard which could prove helpful for the data scientists and the project managers to carry out the above mentioned tasks and then ultimately sell

the AI product with more confidence and clear informative insights to the end-users who would then be able to trust this product.

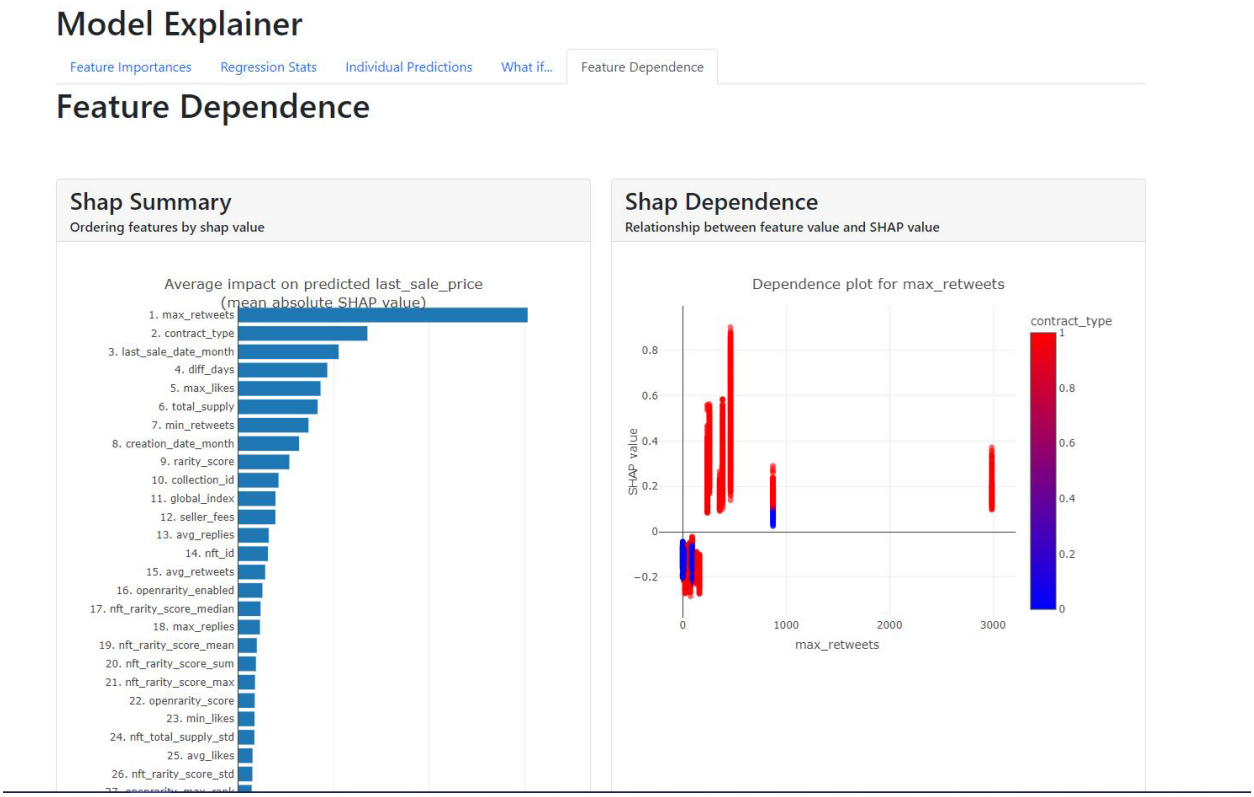
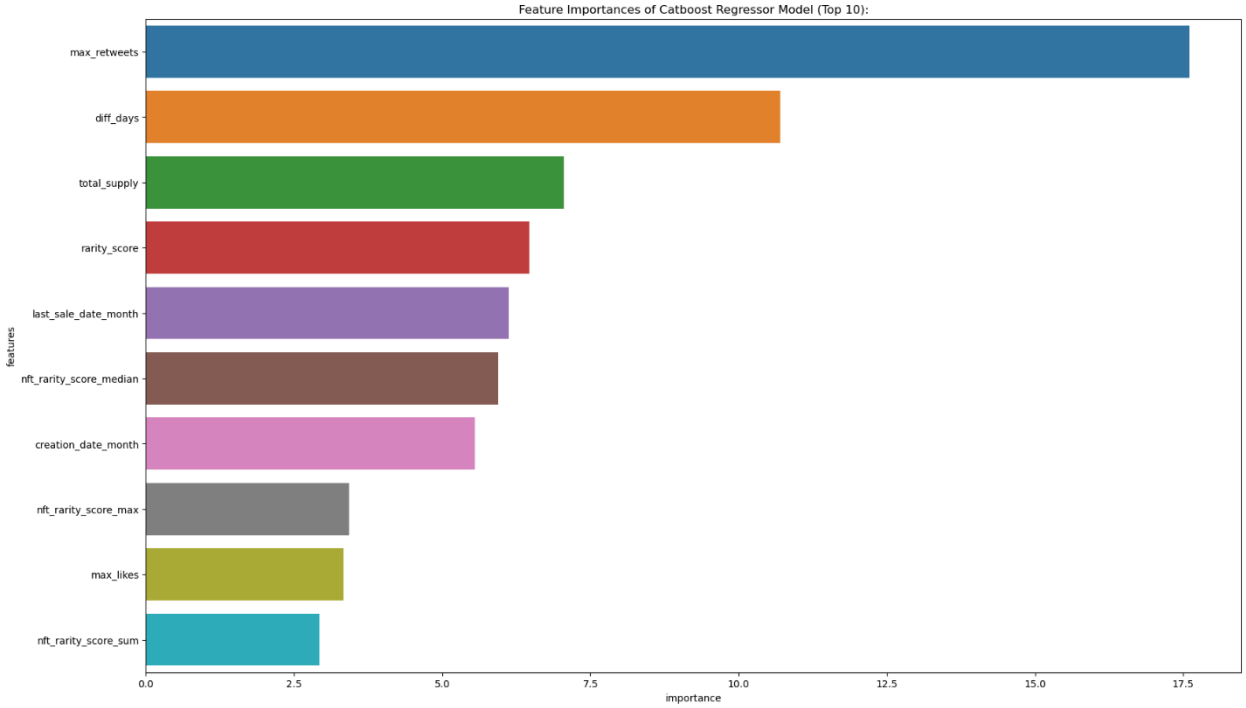


Image 1. Screenshot of Model Explainer Output

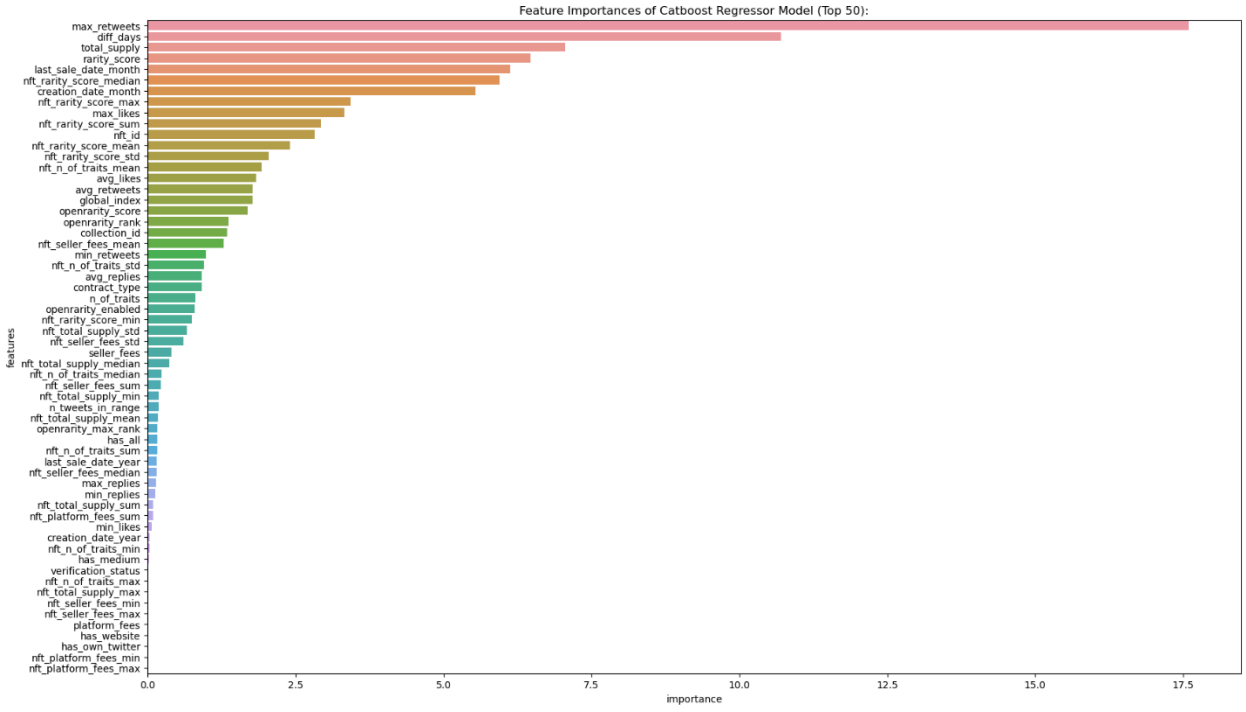
Argument

The main finding of the data interpretation was that some social media characteristics like "max retweets and max likes" as well as other group-based characteristics such as "rarity score, max rarity score, median rarity score" and other features like "difference in days, total supply, last sale date, creation date month" were most helpful in predicting the target variable, the sale price of NFTs. The feature importances plot (Plot 1) produced by the best model from our selected regression approaches can be used to confirm the causality. So, any change in these features mentioned above will affect the price of NFTs in either direction. The plot shown below is the top 10 feature importance plot.



Plot 1. Feature Importances of Catboost Model (Top 10)

The plot shown below is the top-50 feature importance plot.



Plot 2. Feature Importances of Catboost Model (Top 50)

Design Track

Intervention

Our final design is web application for the users to put input data and get prediction on NFTs. The main purpose is to let stakeholders to easily interact and utilize the app to generate precise price predictions for their choice of NFTs to secure healthy returns on their investments. The web-app would also have an additional feature of generating visuals of exploratory analysis as well as complete visualization reports to help the users understand the underlying patterns in the features involved in contributing to the target.

To measure the effects of our intervention, we can monitor the user engagement and perform A/B testing on certain set of changes or addition of new features in respect to a particular hypothesis with correct metrics in place.

The most important thing to measure the effects of our intervention would be to monitor the model's performance in production. If the error rate falls below a certain threshold, it would mean the model's performance deteriorated. We could then regress it back and retrain it with better data and push it back to production ready for use by the users.

Design

As our final design for this project, we created a mock-up of the web application as the end-product for the users to interact with and make use of the application as they see fit. The below images are snippets of web app. For efficient use of the web app, the design is simple with buttons to implement actions needed in straightforward manner. With the ease of use, users can utilize the application without many difficulties and gain wanted information related to NFTs they would like to invest in.

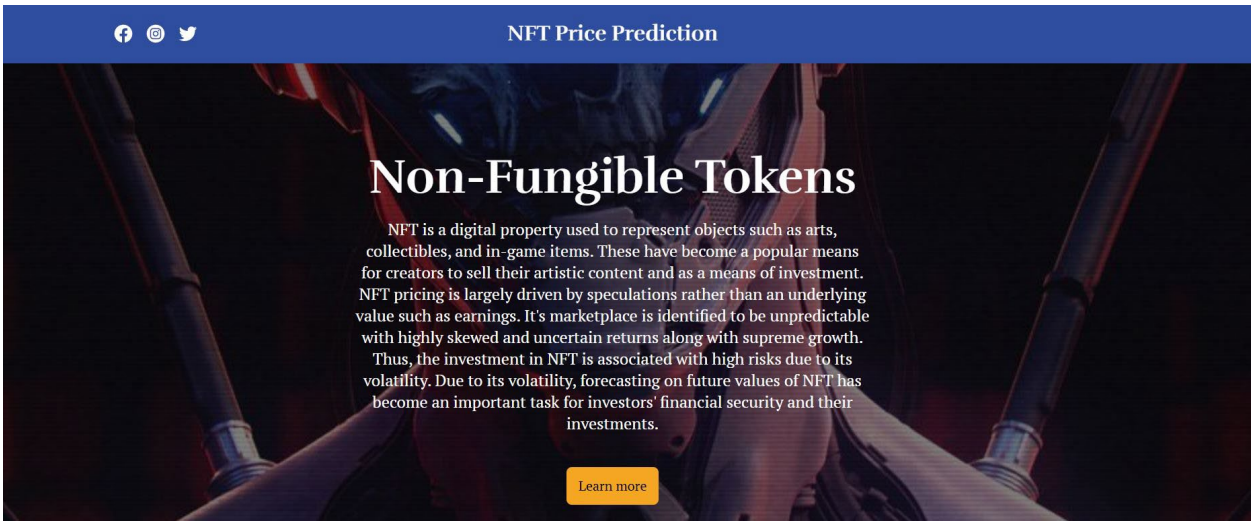


Image 2. Screenshot of Homepage of Web Application

The screenshot above (Image 2) from the homepage of the web-app gives a brief introduction to the application. The concise introduction gives brief overview and information on NFTs to general users, which would help the audience to gain interest and demand for further exploration and explanation on NFTs.

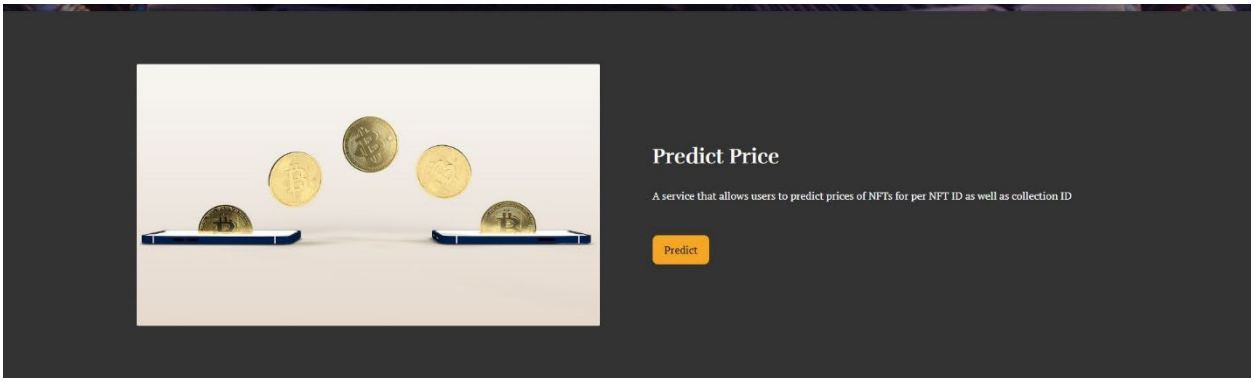


Image 3. Screenshot of Web App on Predict Price Feature

This screenshot (Image 3) above from the homepage shows our first feature of the web-app which is predicting the prices of NFTs. Once the users click on the predict button, it would re-direct them to another webpage where they would see option to input data and then click on predict to generate the price of the particular NFT they wanted to based on their input.

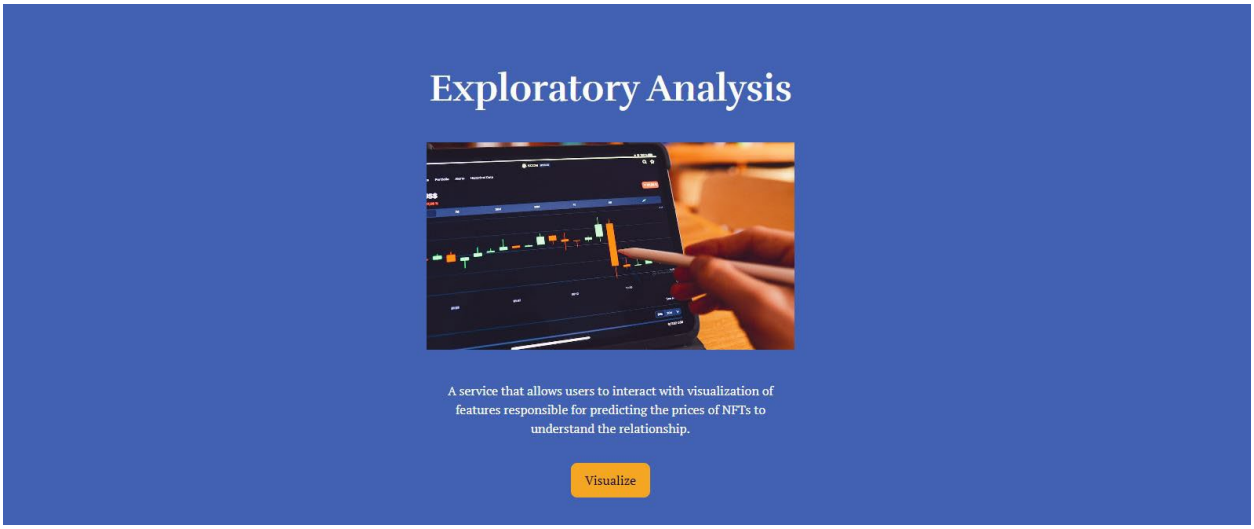


Image 4. Screenshot of Web App on Exploratory Analysis Feature

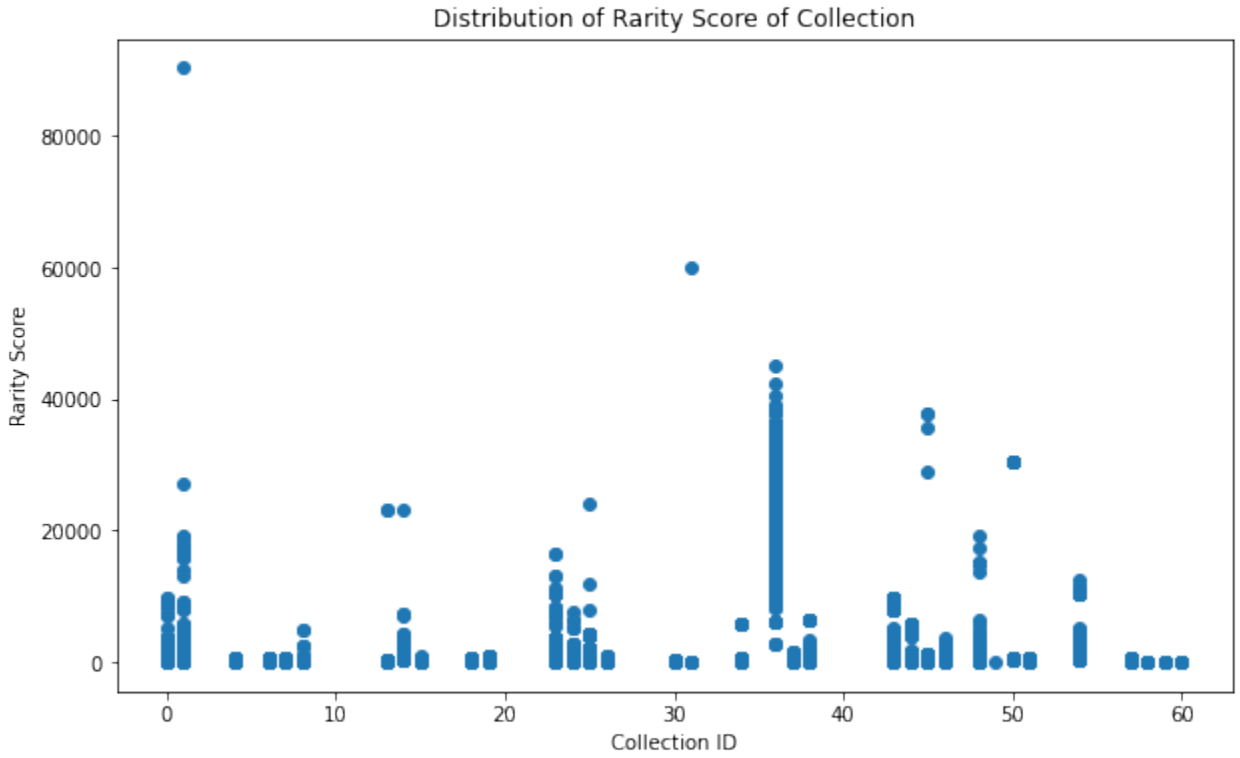
This is the second feature of our web-application which shows the visuals of the features involved in predicting the prices of NFTs (Image 4). This would help the users understand the inter-relationship of the features as well as the relationship of the features with the target variable. Also, the users would get an idea of the underlying patterns within the attributes. In a nutshell, the users would have access to the EDA process of ours to understand the variables involved in price prediction in a better way. All of these would help them make better decisions to secure their investments by maximizing profits.

We haven't considered any alternative design but we are open to adding new features to the web-app or make changes based on the feedbacks we receive from the users as well as the stakeholders.

Data Visualization

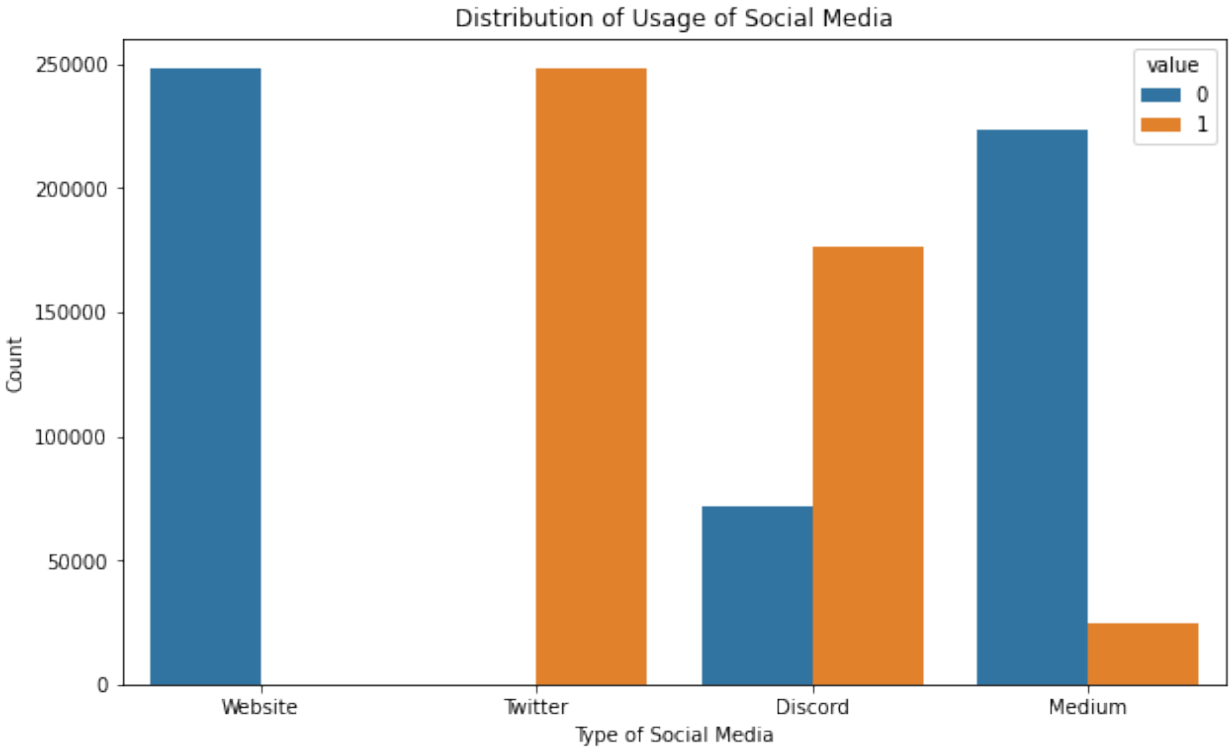
We will be discussing few of the visuals amongst all to showcase how the stakeholders and the audience would be benefitted from the visualizations we have created. Top factors responsible the most for determining the prices of NFTs. The top-10 feature importances plot (Plot 1) helps the users understand what factors contribute most to the determination of prices and then they would be able to do their own additional research and take calculated decisions. From the plot shown below, those features are maximum retweets, difference in days from creation date to last sale date, total supply, rarity score, etc.

Looking at the feature importance plot (Plot 1) above, rarity score is one of the significant attributes that the prediction of sale price of NFTs. Thus, including the EDA plot of rarity score would provide insight to users by showing the trend. The plot attached below tells the users how rare the NFTs are based on their collection ids. Rarer the NFTs are, higher the price would be and vice-versa. With this information, users would be able to understand how our prediction models work and look for appropriate NFTs they would like to invest in accordingly.



Plot 3. Distribution of Rarity Score of Collection

Since one of the critical qualities that helps the prediction to be better is social media related features, it would be important to showcase the EDA on social media attributes. This plot attached below showcases the usage of social media. Not only the group attributes like rarity score and total supply are responsible for predicting prices but also the social media plays a very vital role. So, the users would be able to gain insights about the social media usage for their NFTs. From the plot shown below (Plot 4), Twitter and Discord seem to be the two most important social media platforms which have wide variety of information for the most of the NFTs.



Plot 4. Distribution of Usage of Social Media

Along with visualizations presented above, we will add more information on other EDA results through report obtained from auto EDA. The report will be added as another data visualization to give further information on other attributes used to predict on the sale price of NFTs.

Image 5. Screenshot of Result of Automatic EDA

The data visualizations would captivate audience as the information can be easily conveyed with straightforward plottings. With the thorough analysis of attributes, the users can make better decision in choosing "right" NFT that they would like to invest in. With the easily interpretable plots, our intention to display transparency and effectiveness of our model can certainly be conveyed to our audience. Also, the web application can help users to easily compute prediction price, which would help users to maintain secure investments into NFTs.

Ethics

Working with the NFT dataset can raise several ethical issues that we should consider. Transparency and accessibility are values that are taken into consideration when creating our design. Our web application showcases the data visualization highlighting the underlying patterns. Users can immediately interact with, understand the patterns by gaining crucial insights and make calculated informed decisions. Our design also helps the users predict prices of the NFTs along with the exploratory visuals. The users can make use of these predictions as well as additional information from the visualizations to do their own extra research to make good investments and secure profits.

As a result of our work, potential moral applications include encouraging responsible decision-making in the purchasing and selling of NFTs by fostering transparency and raising market confidence. On the other hand, there are immoral ways to use our data, such as market manipulation. One possibility is for an individual to use the data for their own benefit by spreading false information on social-media platforms, manipulating the sentiment indicator, and taking advantage of the NFT market.

Our design may have an unfavorable impact on underrepresented, marginalized, underserved, and low-resource people. One of the problems causing disproportional effects, which might reduce the advantages of accessing the prediction model and NFT market, can be limited access to the technology.

We need to consider environmental concerns when using NFT data. Energy is needed to manufacture and transfer NFTs during the selling and purchasing processes. We should therefore think about how to address the problem of NFT-related carbon footprints.

People can exhibit their creations in more convenient and open ways. We can encourage sustainable practices in the arts since they will have a better understanding of the market if we provide a forecast model for NFT price. We can stimulate innovation and originality in their respective industries because of improved accessibility and convenience.

References

[1] Ramit Sawhney, Megh Thakkar, Ritesh Soun, Atula Neerkaje, Vasu Sharma, Dipanwita Guhathakurta, and Sudheer Chava. 2022. Tweet Based Reach Aware Temporal Attention Network for NFT Valuation. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6321-6332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[2] https://bitgrit.net/competition/17