Final Project

Julian Carbajal
Kathy Dao
Tiffany Le

Introduction

This data analysis delves into Genshin Impact, an open-world action role-playing game developed by miHoYo and HoYoverse. In Genshin Impact, players assemble a team of four characters, engage in combat, solve puzzles, and explore the expansive world of Teyvat. The game features an elemental system with seven elements, and characters wield elemental powers known as visions. The elemental interactions during battle, coupled with the distinct art style, make Genshin Impact a unique gaming experience.

Among its features, Genshin Impact incorporates a gatcha system, allowing players to obtain new characters and weapons through in-game currency spent on banners. Banners function as loot boxes, providing random rewards, with limited-time characters having boosted drop rates. The game, available for free download, generates revenue through in-game purchases, allowing players to acquire additional in-game currency.

This analysis aims to unravel the mechanics, revenues, and features of Genshin Impact through an examination of three datasets: "Banner Revenue," "Character," and "Weapon." The "Banner Revenue" and "Character" datasets, initially sourced from Kaggle, were supplemented by Tiffany with new characters and features like team flexibility and wish total. The "Weapon" dataset was created to explore another game aspect. Data was gathered from reputable Genshin websites such as Paimon.moe, Genshin.gg, and Gamewith.net.

The "Banner Revenue" dataset comprises 56 rows and 31 features, the "Character" dataset contains 132 rows and 81 features, and the "Weapon" dataset consists of 165 rows and 10 features. Noteworthy is the size of the datasets, particularly the "Banner Revenue" data. Given Genshin Impact's three-year existence and completion status of around 53-60%, this analysis provides insights that may see validation or alteration as the game progresses.

To achieve our analysis goals, we focus on key variables extracted from the three datasets.

Revenue

- **days_since_last_banner_1**: The number of days since the last banner 1.
- **summons_1**: The total number of summons made in banner 1.
- **wish_total_1**: The total number of wishes made in banner 1.
- **days_since_last_banner_2**: The number of days since the last banner 2.
- **summons_2**: The total number of summons made in banner 2.
- **wish_total_2**: The total number of wishes made in banner 2.
- **revenue**: The revenue associated with the banner/update.
- **weapon_1**: weapon type associated with 5 star char
- **5_star_characters_1**: list of 5 star characters (most sought after)
- **banner_days**: length of a specific banner in days.
- **re-run**: whether the banner event came back.
- **wish_total**: The total number of wishes made.

Character Dataset

- **region**: Genshin region where the character originated from.
- **birthday**: Birthday of the character
- **model**: height (tall, medium or short) and gender (male or female) of the character
- **Base HP, Atk, and Def (at max level)**: three different variables that represent the HP, Atk, Def stats of a character at max level.
- **flexibility**: a calculated average of a character's `main_dps_rating`, `sub_dps_rating`, and `support_rating` rating; the value ranges from 0 to 1. This illustrates how flexible a character is in performing in different roles.
- **main_DPS_rating:** : a numeric rating from 1 to 6 that defines the character's effectiveness in a Main DPS (Damage per Second) role. The Main DPS character in the team is the primary "on-field" damage dealer.
- **sub_DPS_rating:** a numeric rating from 1 to 6 that defines the character's effectiveness in a Sub DPS (Damage per Second) role. The Sub DPS character in the team is an "off-field" damage dealer.
- **support_rating:** a numeric rating from 1 to 6 that defines the character's effectiveness in a support role. The team's Support character(s) heals, gives stat boosts, and/or shields their teammates.
- **ascension_specialty:** character stat that is increased when leveling up the character's

main level. The stats that can be increased are: “Anemo/Geo/Electro/Dendro/Hydro/Pyro/Cryo DMG Bonus”, “ATK”, “CRIT RATE”, “CRIT DMG”, “HP”, “Energy Recharge”, “Healing Bonus”, or “Elemental Mastery”

- **vision:** vision/elemental power that the character can use.
- **weapon_type:** the weapon that the character wields
- **rarity:** rarity of the character

Weapon Dataset

- **weapon_type**: type of the weapon
- **secondary_stat**: weapon stat that is increased when leveling up the weapon main level. The stats that can be increased are: “Physical DMG Bonus”, “ATK”, “CRIT RATE”, “CRIT DMG”, “HP”, “Energy Recharge”, “Healing Bonus”, or “Elemental Mastery”
- **rarity**: rarity of the weapon. The weapons can either be of a 3, 4, or 5 star rarity.
- **obtain_method**: binary variable that states if a weapon can be obtained through exploring. 1 if it can be found while exploring and 0 if not.
- **maxed_secondary_stat**: maxed secondary stat when the weapon is fully leveled up.

Question #1: When considering banner revenue, wish total, 5-star character summons, and average days since the last banner (amongst the 5 stars), what clusters emerge, and what characterizes those clusters?

The first question delves into identifying distinct player segments in "Genshin Impact" by analyzing various metrics like banner revenue, wish totals, character summons, and the interval since the last banner. This segmentation aims to reveal underlying patterns in player behavior and spending, which are crucial for personalized marketing, enhancing player experience, and optimizing revenue strategies.

Answer

The implementation of Principal Component Analysis (PCA) strategically reduced the dimensionality of the dataset, thereby streamlining the data architecture while preserving significant variance, crucial for maintaining the integrity of the data's structure. This simplification was paramount for the subsequent clustering process. The process involved in selecting the appropriate number of PCA components was pivotal, ensuring a compromise between data simplification and the retention of substantive variance. Following the PCA application, the K-Means clustering algorithm was utilized to segregate the player base effectively. The utilization of the Elbow Method was crucial for this process, aiding in the ascertainment of an optimal cluster count. This method entails plotting the aggregate of squared distances from each data point to its designated centroid and identifying the 'elbow' where incremental clusters cease to yield a significant diminution in this sum. An in-depth examination further gathered insights into distinct player segments, delineated by the clusters. These segments comprised high-spenders, regular players, and sporadic players, each distinguished by unique behavioral patterns in expenditure, engagement, and game interaction preferences. Evaluative metrics like the silhouette score were instrumental in gauging the delineation and cohesion of these clusters, with a superior silhouette score signifying pronounced definition and distinctiveness of clusters. Various indices such as Calinski-Harabasz and Davies-Bouldin furnished additional validation of the clustering efficacy by quantifying cluster density and separation. The graphical plot indicated that features such as "wish_totals1" and "revenue" were potentially significant contributors to the primary principal component, inferred from their horizontal alignment, thus hinting at their pivotal role in explaining the dataset's variability. Conversely, "days_since_last_banner_1" and "days_since_last_banner_2" appeared to have a more pronounced contribution to the secondary principal component, given their proximity to the vertical axis. The clusters were discernible along the principal component 1 (PC1) axis, signaling that disparities in "wish_totals1" and "revenue" were fundamental in differentiating among the clusters. After interpreting that results it appears that one cluster epitomizes players with high revenue generation and substantial wish totals, whereas another might typify players with lower figures in these aspects, among other possible delineations.

Reasoning

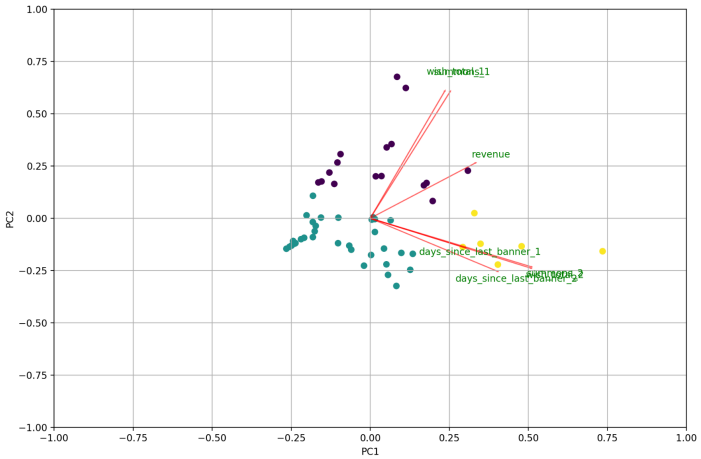

Figure 1A. Bi-plot

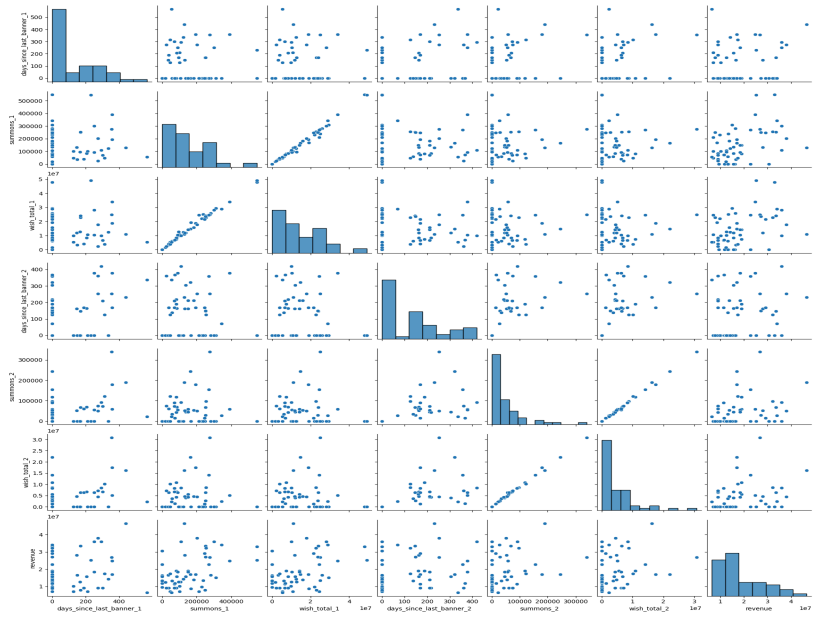

Figure 2A. Pairplot for clustering

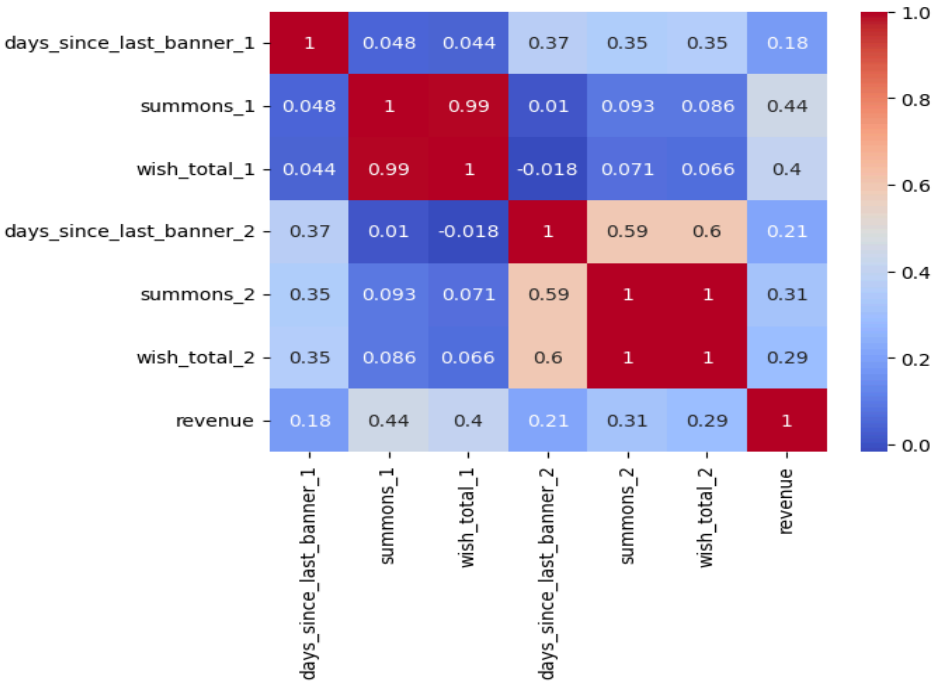

Figure 3A. Heat Map Index for feature importance

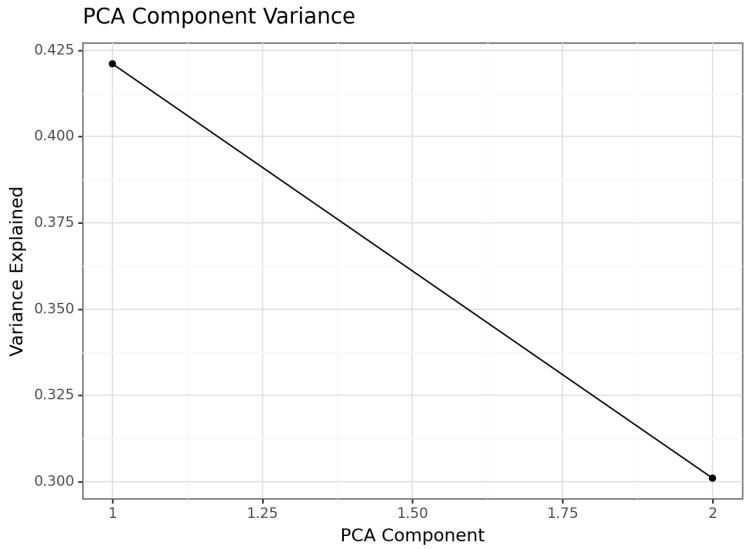

Figure 4A: PCA Variance Explained

Performance metrics	
Inertia:	104.84147767268115
Silhouette Score:	0.4407438501024679

Calinski-Harabasz Index:	45.04800872170341
Davies-Bouldin Index:	0.8113118301067489

Question #2: What is the impact of Banner Duration on Revenue: Does the duration of a banner have a significant impact on its total revenue?

Question Change: I changed question 2 from “Which banners have the highest banner revenue, and which variables (5-star characters, weapon, rerun, mixed, wish total, and duration) have the strongest relationship and impact on the banner revenue?” to “What is the impact of Banner Duration on Revenue: Does the duration of a banner have a significant impact on its total revenue?”. As I was going through the analysis for the first question, my models kept performing very poorly. Despite trying several different model types, the metrics like MSE and MAE were in the trillions, so I went ahead and changed the question to simplify what variables I needed to create a new model. Unfortunately, the model performance was still poor and I got an unreliable model.

The second question aimed to identify the highest revenue-generating banners in "Genshin Impact" and to explore how various factors such as character attributes, weapon stats, and banner durations influence this revenue. Understanding these relationships is pivotal for optimizing future banners and enhancing profitability.

Answer

The analytical approach encompassed a meticulous process of data integration and preparation, where the merging of weapon and revenue datasets became pivotal in crafting a comprehensive view for analysis. Ensuring data integrity and consistency, pre-processing underwent conversion of key variables into numeric formats and handling of missing data to create a robust dataset for modeling. The modeling strategy embraced a multifaceted approach, deploying Linear Regression, Random Forest, and Gradient Boosting Tree models to navigate the analytical journey and compare performance metrics effectively.The in-depth analysis probed into the banner performance, shedding light on banners that garnered

exceptional revenue, thereby unlocking insights into player preferences and spending trends. The investigation into this unveiled critical factors like the total number of wishes, weapon stats, and banner durations, underscoring their substantial influence on revenue streams. These findings suggest that player engagement and the allure of the offerings are instrumental in driving revenue. Model evaluation conducted using quantitative metrics such as Mean Squared Error (MSE) and R-squared values. Regrettably, these metrics painted a stark picture of the model's predictive prowess and accuracy, or lack thereof. With exorbitantly high MSEs and negative R-squared values, signaled the reality that the model was struggling to capture the nuanced relationship between independent variables and the dependent revenue variable. Concluding the analysis, the clustering elucidated distinct player segments, each with its own defining traits, which could be leveraged for targeted marketing and engagement strategies. Predictive modeling shed light on the primary factors propelling banner revenue, delivering strategic insights for future banner development and optimization. Despite the model performance being lackluster, finding that wishs_total was the lead contributor to revenue is not a bad feature to essentially invest in. As a recap, wishes in the game are like lottery tickets. Players can accumulate them for free by playing the game and progressing through the story, but the game also has a marketplace where you can buy wishes to increase your chances of winning a sought-after 5-star character. So the end result of wish_totals being of primary importance is not far fetched in spite of the poor data modeling performance.

Reasoning

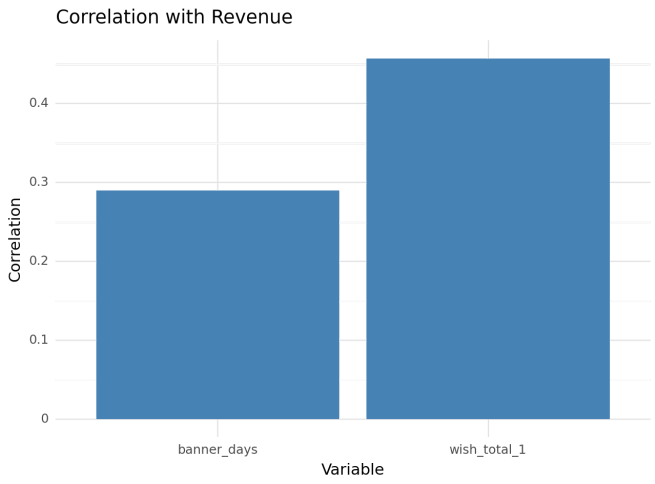

Figure 1B. Correlation between Revenue

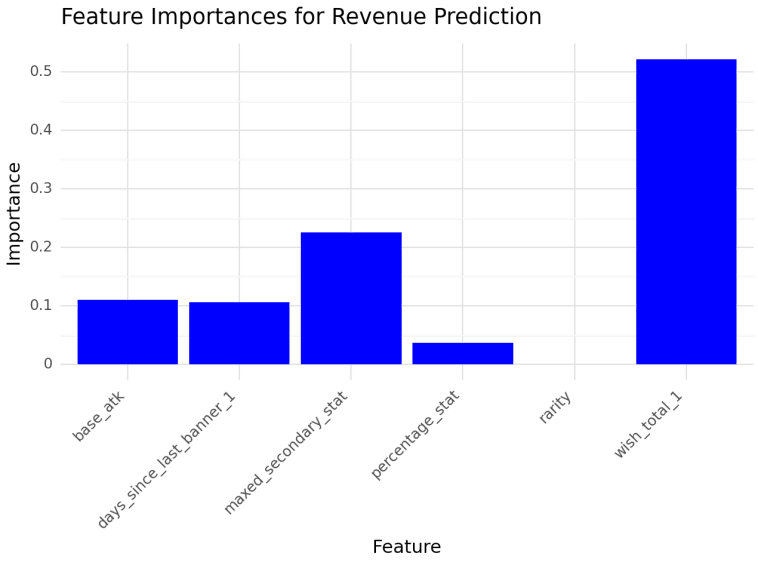

Figure 2B. Feature Importance for Revenue Plot

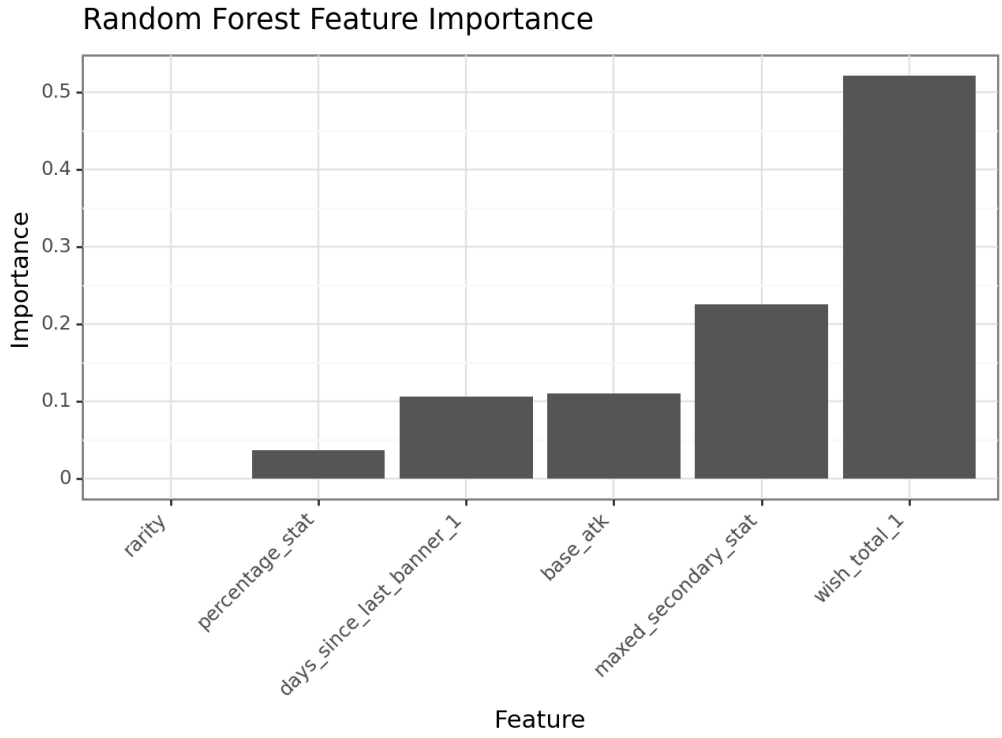

Figure 3B. RF Feature Importance

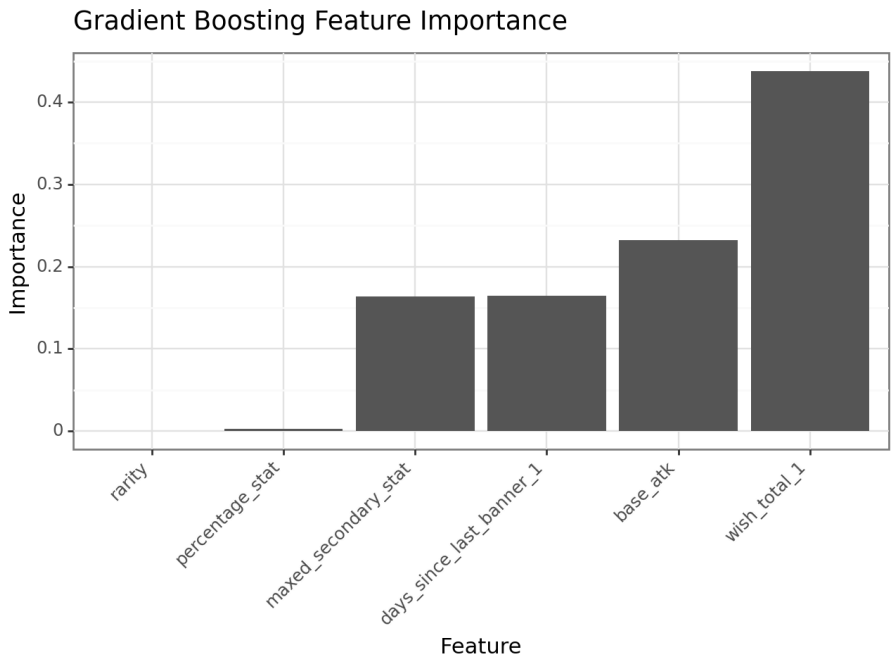

Figure 4B: GB Feature Importance

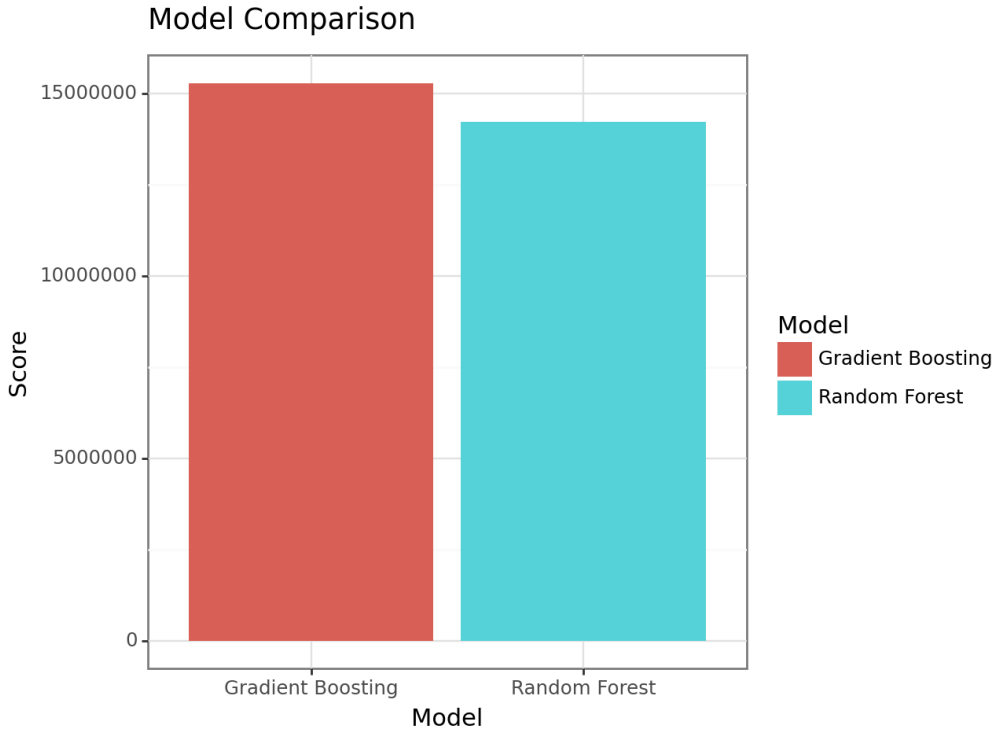

Figure 5B. Model Comparison

Question #3: When predicting a Character’s Rarity, which predictor “Lore-Related” features or their “Playstyle Stats” improves the Accuracy the most compared to a model with all other variables except itself?

Question Change

The original question to be analyzed was, “Can we predict a character's talent book based on their team role, birthday, region (e.g., Mondstadt, Liyue), and vision? Additionally, what features contribute most to the model's accuracy?” While analyzing this question, I noticed that the dependent categorical variable had many categories. As mentioned above, the dataset is on the smaller end, so having one category be shared amongst only 2-4 characters can make it difficult for the model to predict or avoid overfitting. That being said, the question was simplified to do a binary classification of whether a character is a four (4) or five (5) star rarity.

Answer to the Question

The analysis of the "Character" dataset employed three classification models—Logistic Regression, Random Forest Classifier, and Gradient Boosting Classifier—to explore the impact of "Lore-Related" and "Playstyle Stats" features on character rarity classification. Among these, the "Playstyle Stats" features demonstrated the most significant enhancement in accuracy compared to a model incorporating all other variables except itself.

Reasoning

In order to find our answer, we tried three different approaches. We put our data and its features through three different models to see which one is the best at predicting or figuring out the character’s rarity. A new feature, "Zodiac Sign," was introduced to enrich the dataset, providing additional character information based on their birthdays.

Logistic Regression Model

Ridge penalty and K-Fold cross-validation were applied to the Logistic Regression models to mitigate overfitting and enhance performance on unseen data.

Random Forest Classifier Model

Hyperparameter tuning was performed on the Random Forest Classifier models to optimize inputs and settings, addressing overfitting concerns and regulating model performance on both training and testing data.

Gradient Boosting Classifier Model

Given its suitability for smaller datasets, the Gradient Boosting Classifier model was chosen, and hyperparameter tuning was conducted to refine its performance.

Assessing the Performance of Each Model

Performance metrics such as Accuracy, Precision, Recall, F1 Score, and ROC Auc were used to evaluate the nine models across training and testing sets.

Results

After conducting the performance analysis on the three models for the two subset dataset, the Gradient Boosting Classifier model scores were significantly higher than the other two models. The performance metrics for each model's training and testing sets are displayed in the table and figure below.

After narrowing down the model whose results will be compared to the rest, we can see, in Figure 1-4C, that while "Lore-Related" features contribute valuable information towards the classification, the model's accuracy significantly improves when considering "Playstyle Stats" alone. The high accuracy of the "Playstyle Stat" model indicates the features strong influence in predicting a Character's rarity.

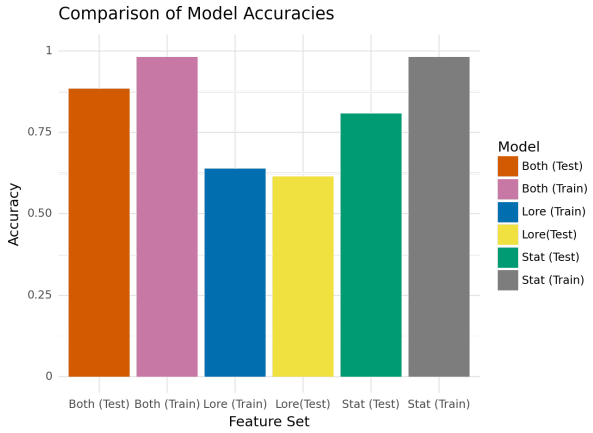

Figure 1C: This bar graph shows the accuracy of a GBC model on the three (3) dataset.

Model	Accuracy
Lore (Train)	0.638095
Lore(Test)	0.615385
Stat (Train)	0.981132
Stat (Test)	0.807692
Both (Train)	0.981132
Both (Test)	0.884615

Figure 2C: This table shows the numerical accuracy scores of each model on the training and testing data.

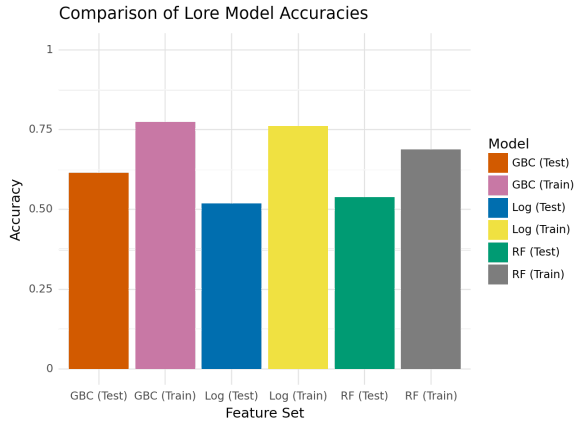

Figure 3C: This bar graph shows the accuracy of each classification (GBC, Logistic, Random Forest) on the "Lore-Related" features dataset.

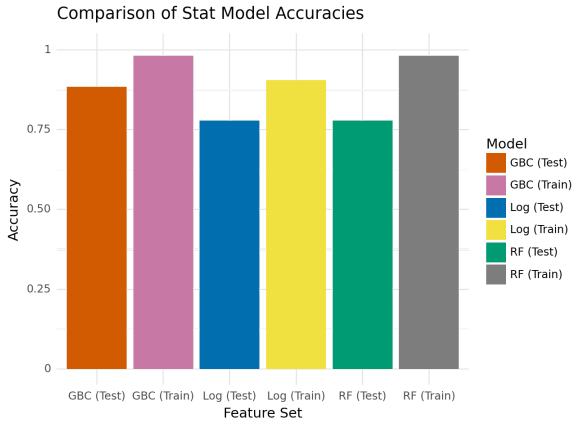

Figure 4C: This bar graph shows the accuracy of each classification (GBC, Logistic, Random Forest) on the "Playstyle Stat" features dataset

Feature Importance

After conducting the classification of the data from each model, we calculated the importance of each feature in the combined dataset model. Upon analysis, we found that the top ten features contributing to the model's accuracy are the ones listed in **Figure #5C.** To enhance clarity, we calculated the collective feature importance for the two feature categories (Lore and Playstyle) to see which category held more importance to determining the character's rarity. While the "Lore-Related" features collectively scored 0.124135, the "Playstyle Stats" features had a significantly higher importance score of 0.758854.

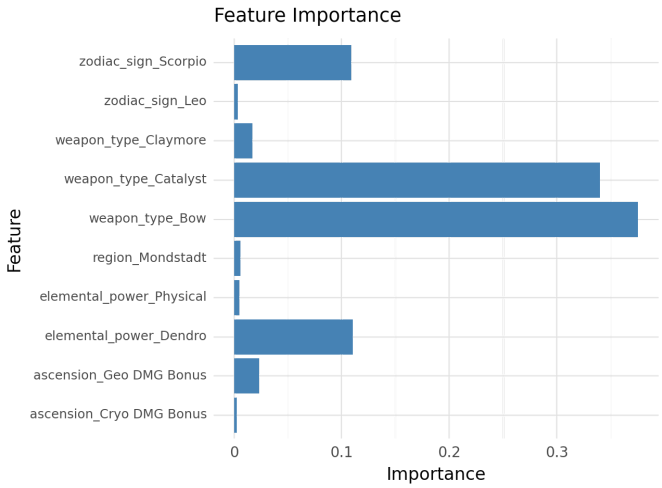

Figure 5C: This horizontal bar graph shows the "Top Ten" features in the entire analysis and how much each feature contributes to classifying a character's rarity.

Discussion and Conclusion

After looking at and comparing the accuracy results, we discovered that features related to how characters fight and their abilities (“Playstyle Stats”) are the most important in determining rarity. Understanding this helps game developers and content creators (mainly YouTubers and Twitch streamers), and it guides them [game developers] in creating characters that players will find exciting and valuable. For content creators, it guides them on creating content that their players will love such as focusing on characters that bring a new feature to the game (aka “introducing a new meta” as some gamers would say). So, this analysis helps us understand what makes a character special in Genshin Impact. While a portion of a character’s lore is important, the way they fight and their unique abilities play a big role in how rare and valuable they are in the game. This information is important for the developers at HoyoVerse and for those who make content (YouTube and Twitch streamers) about it, helping them make decisions and guides that players will love.

Question #4: Can we predict the weapon's base ATK based on their weapon type, secondary stat, rarity, obtain method, and maxed secondary stats? Which of the variables has the strongest relationship with the base ATK?

Question Change

The original question proposed to be analyzed was, “Can clustering techniques reveal groups of weapons with similar characteristics regarding base ATK, weapon_type, secondary stats, and maxed secondary stats? How do these clusters align with weapon rarity and obtain methods (purely gatcha (luck) or exploration is involved)?” While analyzing the other questions, I wanted to focus more on predicting values rather than clustering. The other questions centralize clustering, so I wanted to focus more of my effort on data prediction and Supervised Machine Learning.

Answer

Predicting the base ATK of a Genshin Impact weapon, with the variables listed, is feasible, and the Gradient Boosting Regressor model proves to be highly effective in achieving accurate predictions with minimal error. Among the variables considered, the weapon's rarity stands out as the most influential, exhibiting the strongest relationship with the base ATK.

Reasoning

The "Weapon" dataset underwent analysis using three regression models—LASSO Regression, Random Forest Regressor, and Gradient Boosting Regressor—to compare their prediction performances.

LASSO Regression Model

To address overfitting observed in the original Linear Regression model, a LassoCV model was employed, incorporating a penalty to reduce complexity while maintaining accuracy. K-Fold cross-validation was further applied to enhance performance on unseen data. In simple terms, these methods are utilized to improve the model's prediction on data is hasn't been seen before.

Random Forest Regressor Model

Hyperparameter tuning was conducted to optimize the Random Forest Regressor model's inputs and settings, mitigating overfitting (performing poorly on data it has not seen) and regulating performance on both training and testing sets.

Gradient Boosting Regressor Model

Given its suitability for smaller datasets, the Gradient Boosting Regressor model was selected and underwent hyperparameter tuning for performance enhancement.

Assessing the Performance of Each Model:

Performance evaluation involved metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R2 on both training and testing sets. These metrics are used to see how well each model is doing with its predictions.

Results

The Gradient Boosting Regressor model outperformed the other models significantly, as depicted in Figure #1D. Additionally, we can see in Figure #2D that the model's predictions aligned closely with actual values, showcasing superior accuracy.

	MSE	MAE	MAPE	R2
LassoCV (Train)	646.49	20.04	646.49	0.90
LassoCV (Test)	1067.37	25.27	1067.37	0.80

Random Forest Regressor (Train)	1272.61	27.40	0.05	0.79
Random Forest Regressor (Test)	1788.01	32.76	0.06	0.70
Gradient Boosting Regressor (Train)	69.67	3.35	0.006	0.99
Gradient Boosting Regressor (Test)	511.24	11.5	0.02	0.90

Figure 1D: This table shows the performance metrics of each model prediction on the training and testing set.

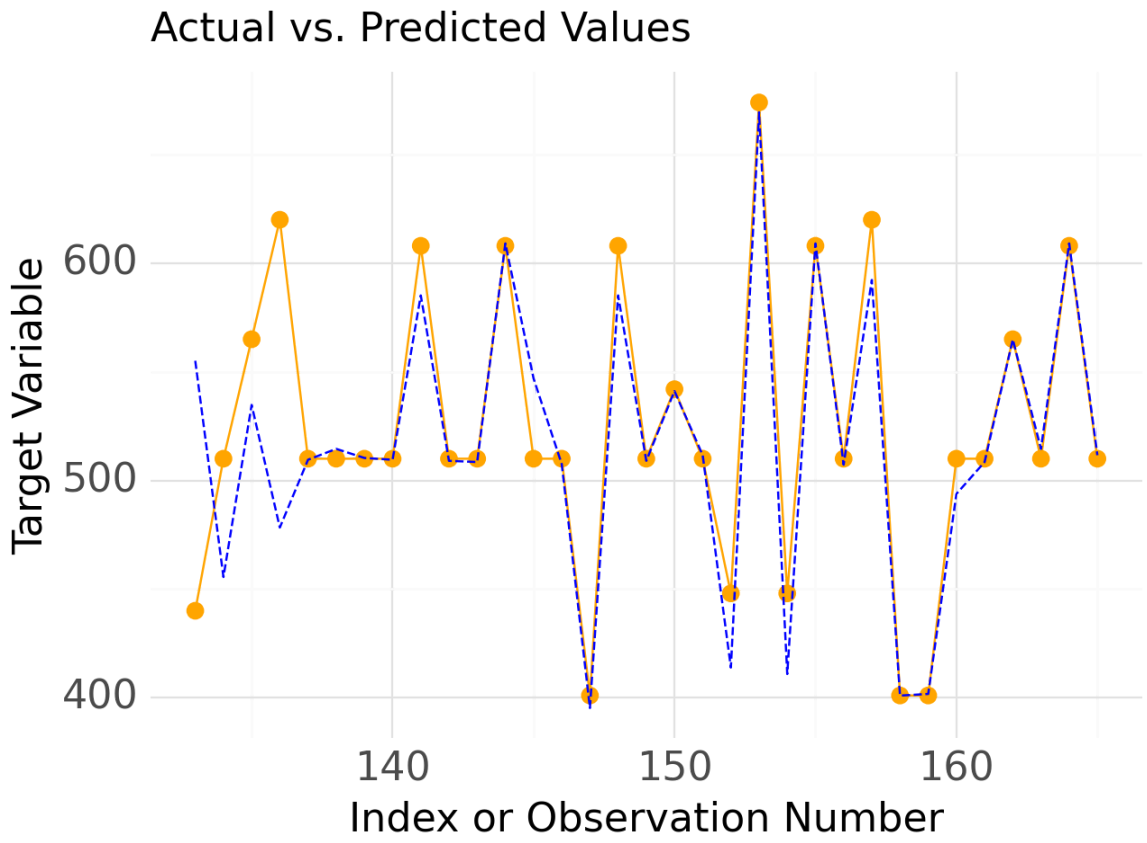

Figure 2D: This line plot shows the actual (orange line) and the Gradient Boosting Regressor's predicted value (dashed blue line).

Feature Importance

We also found that some features are most important in making a prediction on a weapon's base ATK. The rarity of the weapon turned out to be the most important, followed by the other variables. The importance of each feature is listed below in Figures 3D and 4D.

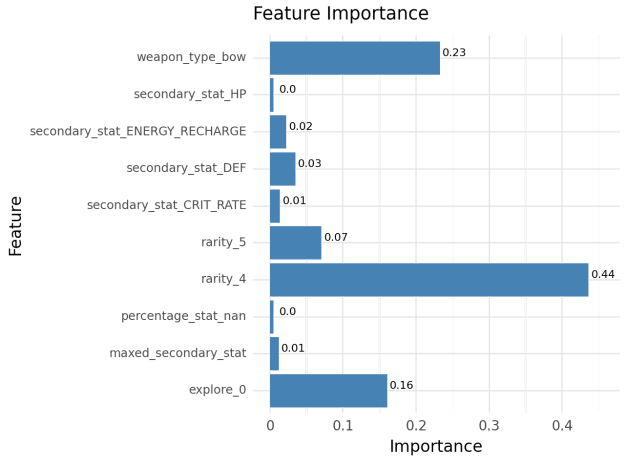

Figure 3D: This horizontal bar graph shows the top ten features that impact the results of the prediction.

Feature	Importance
Weapon Type	0.23455
Secondary Stat	0.078624
Rarity	**0.506313**
Obtain Method	0.161389
Maxed Secondary Stat	0.012113

Figure 4D: This table shows the collective importance of each feature when predicting weapon base ATK.

Discussion and Conclusion

In summary, our analysis of predicting a weapon's base ATK led us to conclude that a Gradient Boosting Regressor model was the optimal choice. The model proved to perform the rest on this data compared to the other two models. Notably, the analysis of feature importance highlighted that the weapon's rarity plays a pivotal role in predicting a weapon's base ATK. With a feature importance score of 0.506, rarity emerged as the most influential variable in this prediction task. This insight simplifies the prediction process, showcasing the significance of considering a weapon's rarity when estimating its base ATK. This finding holds significant implications for game developers and content creators in the Genshin Impact community. Understanding the importance of rarity in predicting base ATK informs decisions related to weapon development and their integration into the open-world combat system. These insights not only enhance the predictive accuracy but also provide valuable guidance for content creators aiming to optimize their engagement on platforms such as Twitch and YouTube through weapon and team guides tailored to user preferences.

Question #5: Can we identify distinct groups or patterns among characters based on their predictor profiles, and do these groups exhibit different levels of flexibility?

Question Change

I changed the question, “When predicting a character’s vision, which predictor (region, weapon type, flexibility, and ascension material) improves the Accuracy the most when compared to a model with all other variables except itself?” to “Can we identify distinct groups or patterns among characters based on their predictor profiles, and do these groups exhibit different levels of flexibility?” because my results identified the groups and patterns among the characteristics and their correlation to their flexibility.

Answer

First, I prepared the data for analysis by ensuring the dataset is clean and contains the necessary variables. Then, Principal Component Analysis (PCA) is strategically integrated to reduce dimensionality, ensuring the resultant clusters are not just statistically significant but also interpretable and resourceful. After, I created a scree plot analysis, emphasizing the "elbow" point, which serves as a pivotal step in determining the optimal number of clusters (K). I also made a cumulative variance plot that aids in deciding how many clusters to create, aligning with the inherent structure of the data. The utilization of the K-means clustering algorithm further enhances the analysis, pinpointing natural groupings or clusters where points within the same cluster share greater similarity than those in other clusters. This methodological choice, coupled with the subsequent analysis of each cluster's characteristics, offers insights into the relationship between predictor variables and character flexibility. Visualizing clusters in a reduced-dimensional space allows for the identification of patterns or trends within the data, contributing to a nuanced understanding of how specific features influence the separation of data points into distinct groups. The comprehensive report detailing these analyses provides a deeper understanding of the interplay between character profiles and flexibility. As a result, I can see that *main_DPS_rating* contributes the most to a character’s *flexibility.*

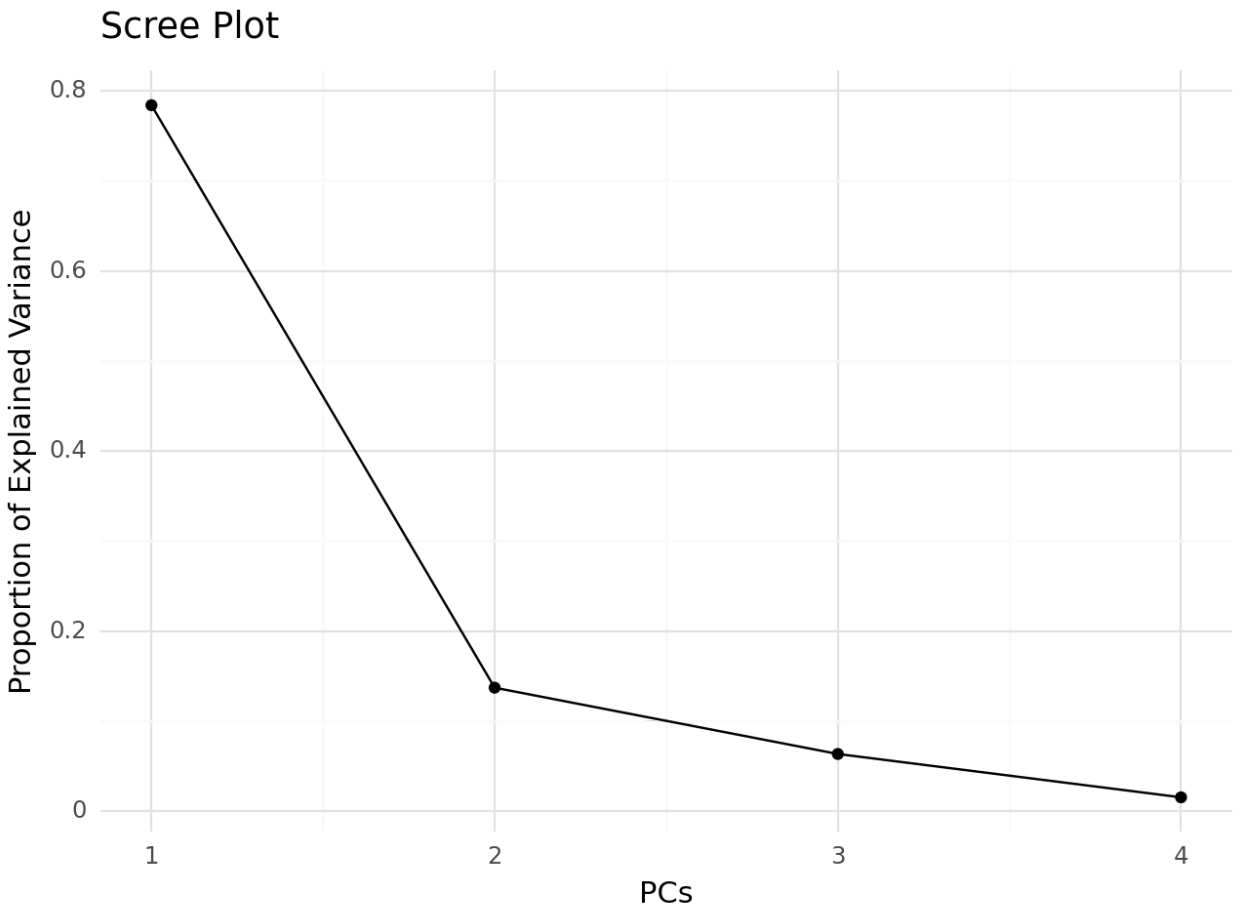

Figure 1: Scree plot to find optimal K

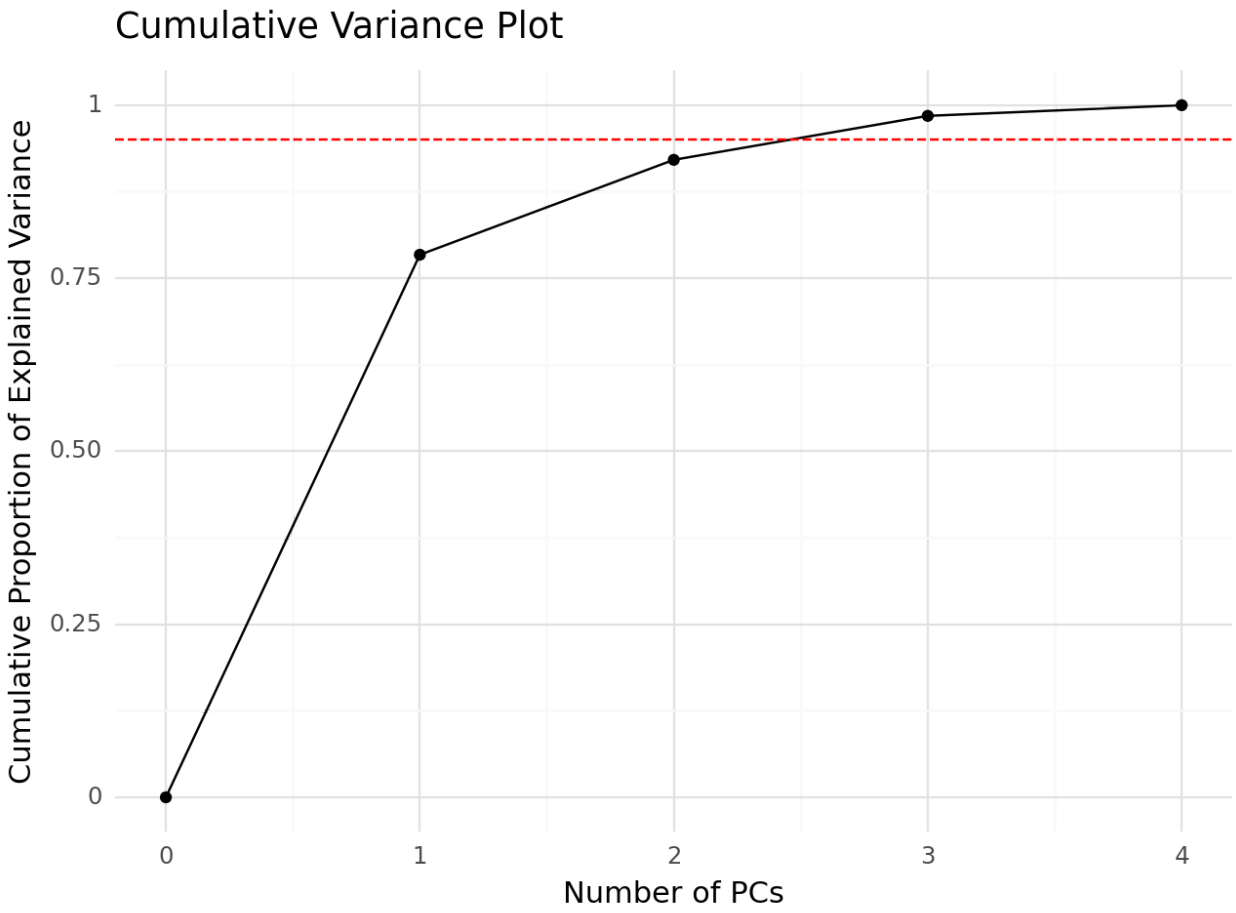

Figure 2: Cumulative variance plot

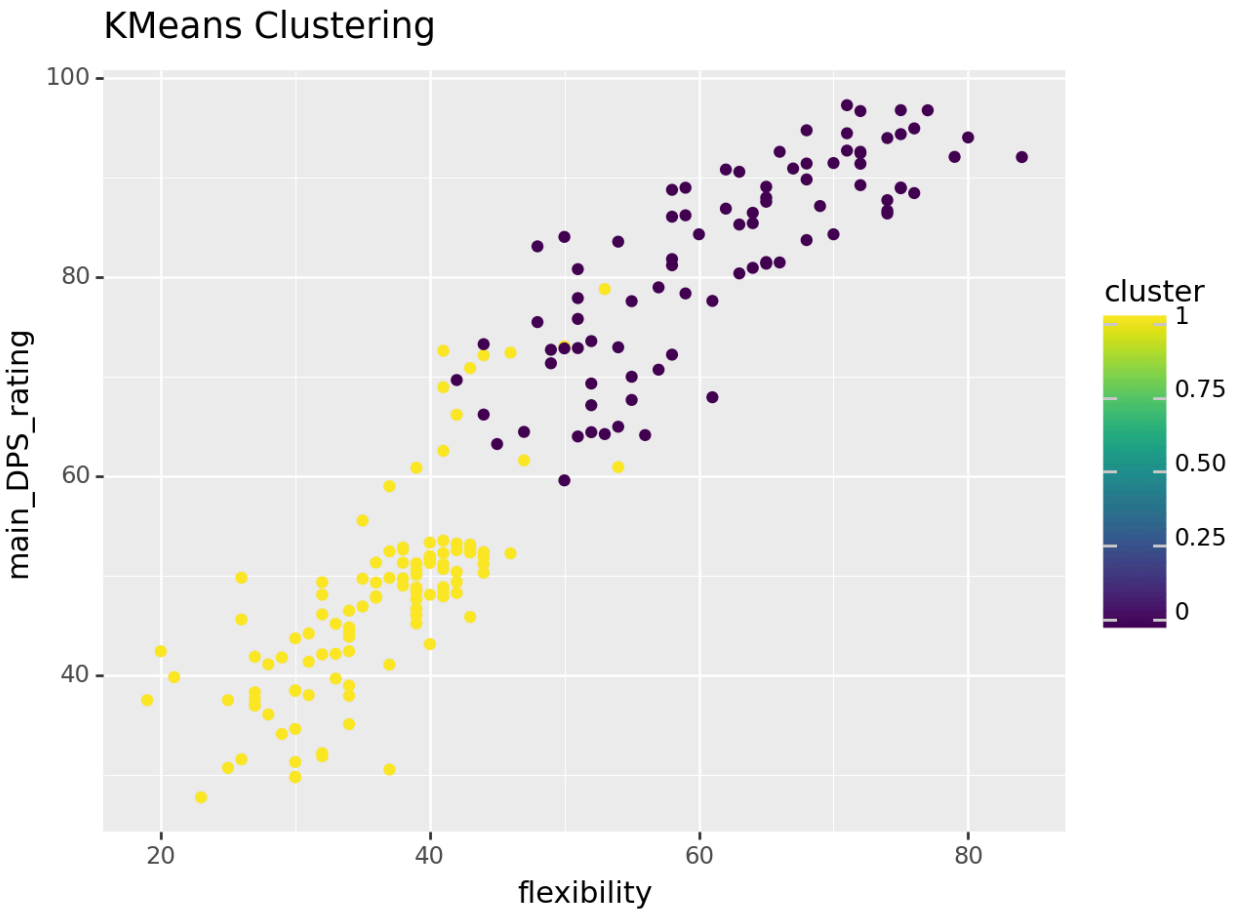

Figure 3: K Means Clustering of main_DPS_rating and flexibility

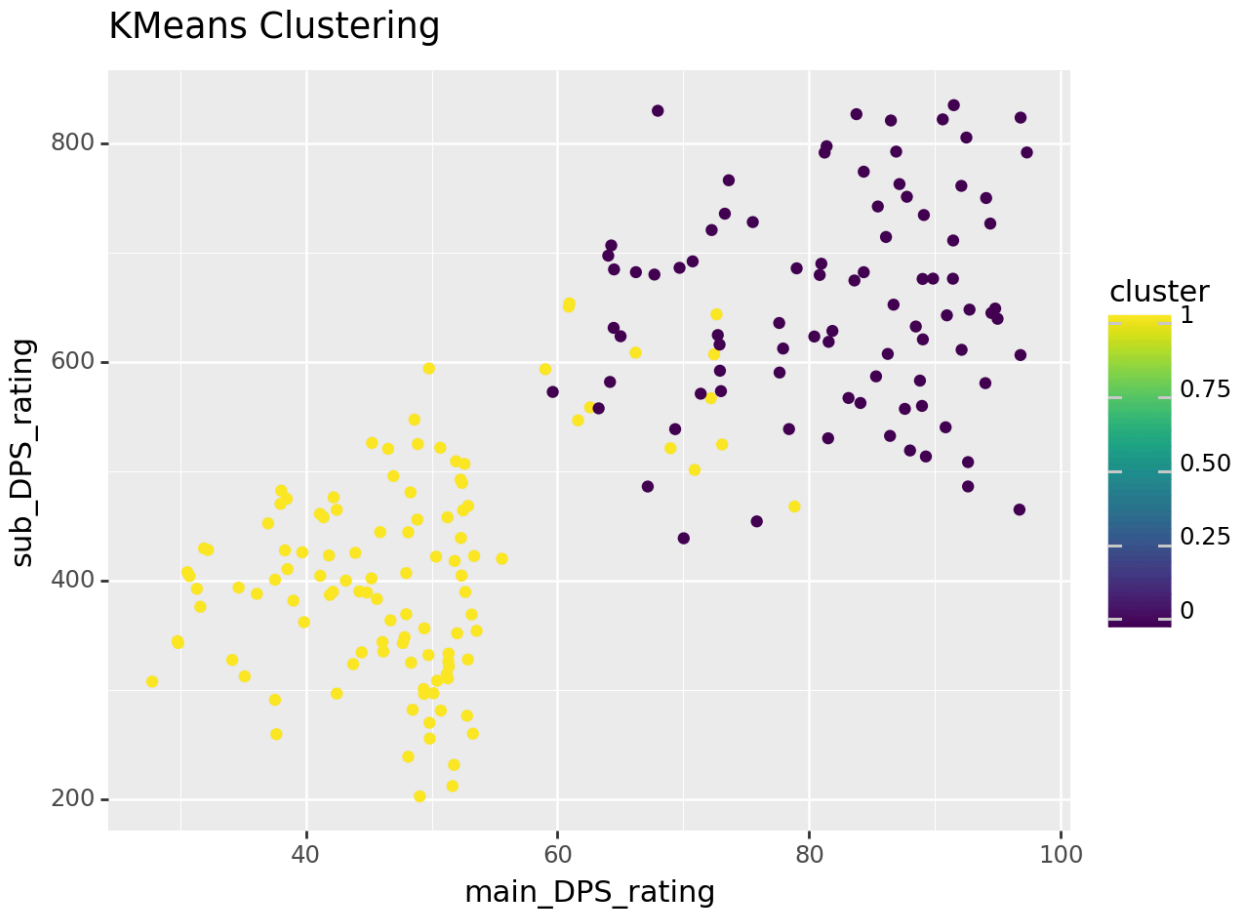
KMeans Clustering
800
600
400
200
sub_DPS_rating
40
60
80
100
main_DPS_rating
cluster
1
0.75
0.50
0.25
0

Figure 4: K Means clustering of sub_DPS_rating and main_DPS_rating

Figure 5: K Means clustering of supporting_rating and sub_DPS_rating

Question #6: Using a supervised model, can we predict a character's team flexibility based on their stats at different ascension levels, vision, and weapon types? What features contribute most to the model's accuracy?

Answer

This analysis aims to predict a character's team flexibility using a supervised model based on ascension_specialty, vision, and weapon_type, with a focus on identifying the features contributing most to the model's accuracy. The initial step involves data cleaning by dropping null values. A logistic regression model is then applied, yielding the following results on both the training and test sets: Accuracy (Training Set): 0.728, Recall: 0.179, ROC AUC: 0.695;

Accuracy (Test Set): 0.722, Recall: 0.181, ROC AUC: 0.701. Subsequent to the model evaluation, a calibration curve is generated, displaying a line aligned with the linear line, indicating good calibration. Feature importance analysis unveils key contributors, with weapon_type (0.630) and rarity (0.407) being the most influential, followed by specific main_DPS_rating categories and ascension_specialty variations. This comprehensive analysis, culminating in a feature importance graph, provides valuable insights into the predictive capabilities of the logistic regression model and the nuanced influences of different character attributes on team flexibility. The result says that weapon type is the main contributor to the character's flexibility.

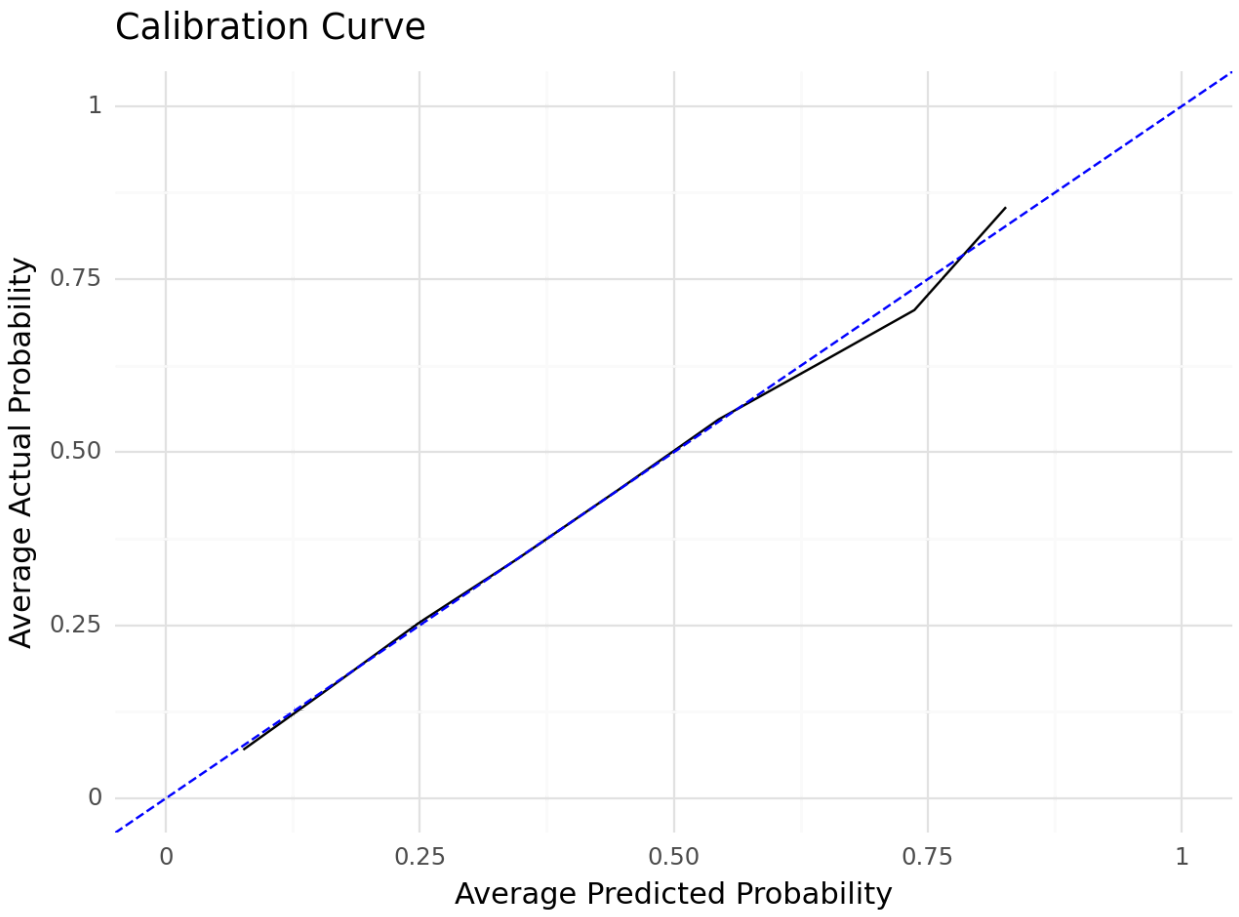

Figure 1: Calibration curve

Logistic Regression Metrics (Training Set)	
Accuracy	0.7284689133887479
Recall	0.17885551948051948

ROC AUC	0.6951613864620101
Logistic Regression Metrics (Test Set)	
Accuracy	0.7215440792905582
Recall	0.1807909604519774
ROC AUC	0.7010029963665447
MSE	0.27845592070944186
MAE	0.27845592070944186

Figure 2: Logistic Regression metrics

Variable	Feature Importance
weapon_type	0.629699
rarity	0.407155
main_DPS_rating_P	0.265181
main_DPS_rating_B	0.253485
ascension_specialty_man	0.232406
ascension_specialty_nonbinary	0.230647
ascension_specialty_other	0.200226
main_DPS_rating_A	0.176087
vision	0.051918
sub_DPS_rating	0.036026
ascension_specialty_woman	0.031474
support_rating	0.003157

Figure 3: Feature Importance table

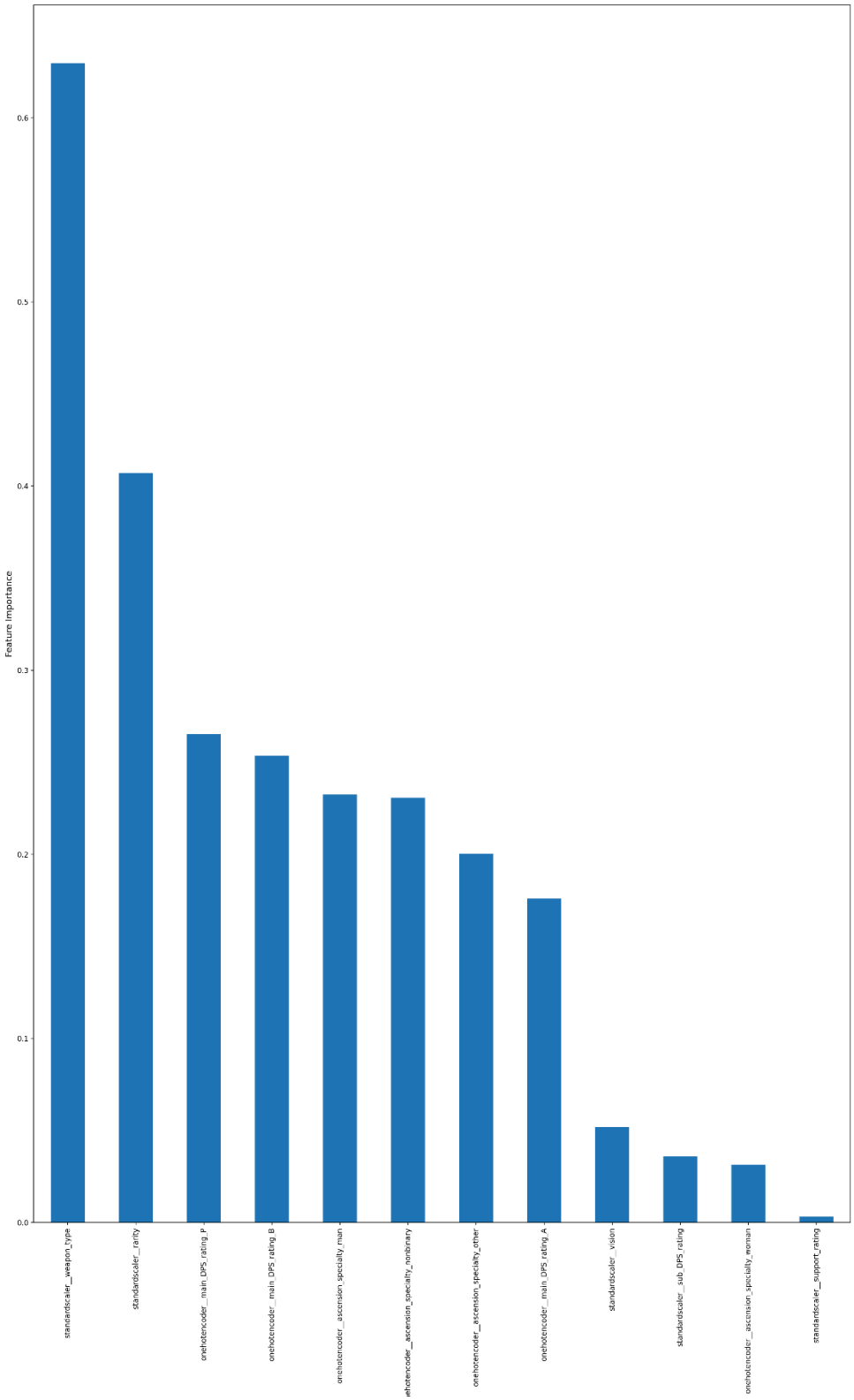

Figure 4: Feature importance bar graph