eBay Auction Competitiveness Analysis

RSM8413 Group Assignment 1

Team 12

[Harrison Nan Lou, Lamana Mulaffer, Peter Salame, Yaojie Cui, Zimeng Li, Zoe Zong]

October 5, 2025

Executive Summary

Project Objective

This project developed predictive models to classify eBay auctions as **competitive or non-competitive** using data from May–June 2004. By identifying auctions likely to attract low competition, eBay can provide sellers with actionable recommendations—adjusting pricing, auction duration, or listing strategy—to increase competitiveness. The goal is to **boost revenue, enhance seller success, improve buyer engagement, and strengthen overall marketplace health**

Key Insights from the Data

- **Auction Drivers:** Low starting prices, categories such as Photography, Electronics, SportingGoods, Monday auctions, and extreme SellerRatings (very low or very high) are associated with higher competitiveness.
- **Dataset Overview:** 1,972 auctions, 8 features, no missing values; 54% competitive. Data was preprocessed with one-hot encoding for categorical variables, scaling for numerical features, and an 80/20 stratified train-test split.

Modeling Outcomes

- **K-Nearest Neighbors (KNN):** Achieves up to 77% accuracy with all features; drops to 69% accuracy for live predictions without ClosePrice.
- **Decision Trees (DT):** Achieve ~81% accuracy with all features and ~72% accuracy for live predictions using top 2 features (excluding ClosePrice).
- **Model Insights:** Decision Trees are more robust for both live predictions (while auction is in-progress) and also post-auction predictions as well.

Business Implications & Next Steps

- Predictive models can **guide sellers to improve auction competitiveness**, optimize listings, enhance engagement, and boost overall marketplace health.
- Implement predictive scoring on live auctions to **proactively flag low-competition listings**, provide targeted seller recommendations, and continuously monitor performance to refine models over time.

Exploratory Data Analysis (EDA) & Data Preprocessing

Approach

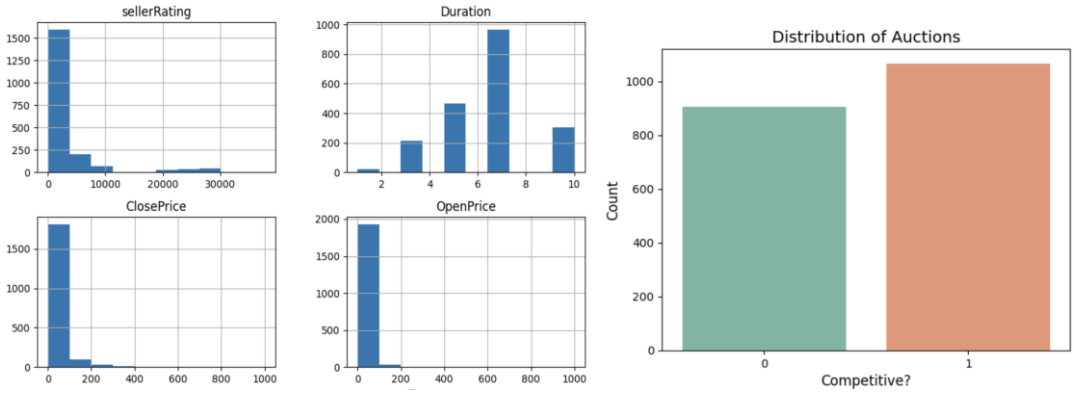
We performed exploratory data analysis (EDA) to understand the dataset's structure, distributions, and relationships, including univariate, bivariate, and multivariate analyses. Data preprocessing included handling categorical variables with **one-hot encoding** (keeping top categories and binning others as "Other"), **train-test splitting** with stratified sampling (80/20), and **scaling** numerical features using the **train data statistics**.

Key Findings

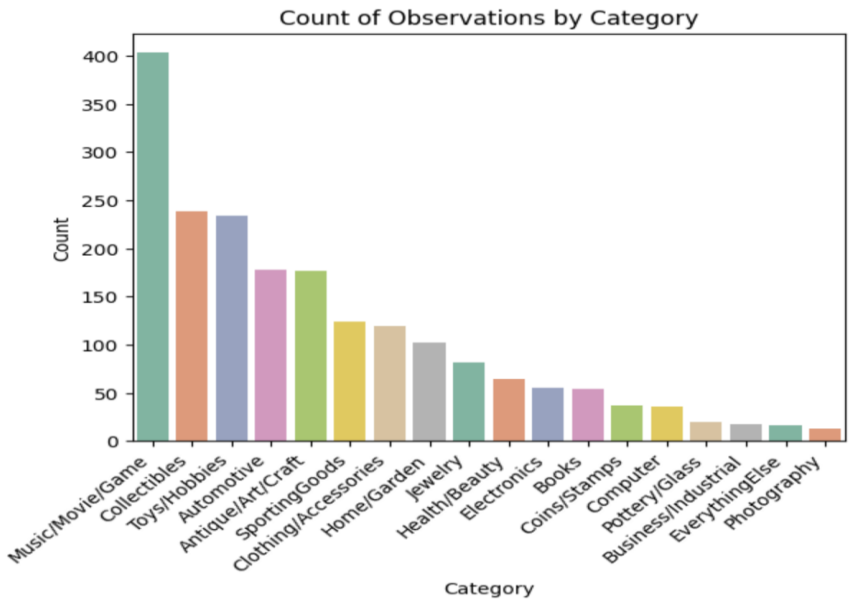
Dataset Overview

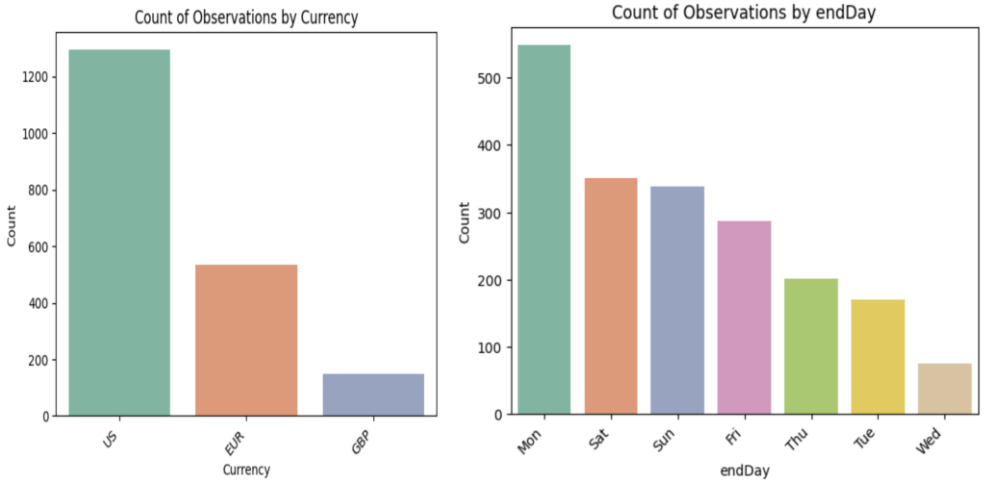
The dataset contains **1972 observations** and **8 variables**, including **4 categorical variables** (Category, Currency, endDay, and Competitive? (as the target)) and **4 numerical variables** (sellerRating, Duration, OpenPrice, and ClosePrice), with **no missing values.**

Univariate Analysis:



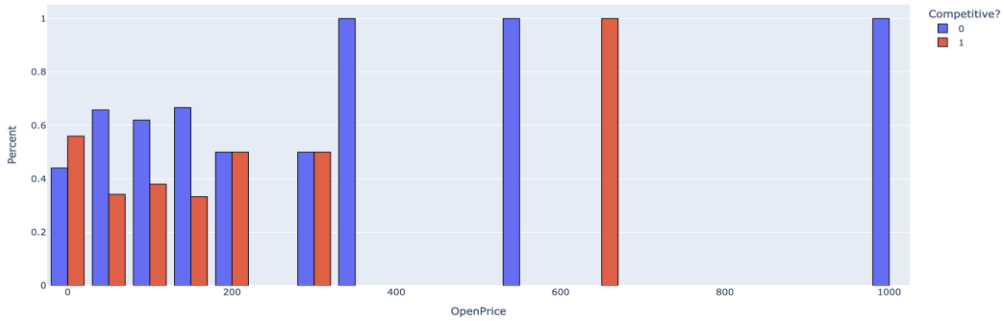
- The target variable is moderately balanced, with 54% of auctions classified as competitive (see: graph above right).
- Numerical features such as *sellerRating*, *OpenPrice*, and *ClosePrice* are highly right-skewed with notable outliers, while *Duration* is dominated by 7-day listings. These patterns indicate the need for potential log transformation or robust scaling before modeling (see: graph above left).
- Categorical variables show strong imbalance—*Music/Movie/Game* dominates the listings, with few auctions in other categories. Most transactions use USD, followed by EUR and GBP. Auctions most frequently end on Mondays and weekends, while midweek (especially Wednesday) is least common. (see: graphs below)

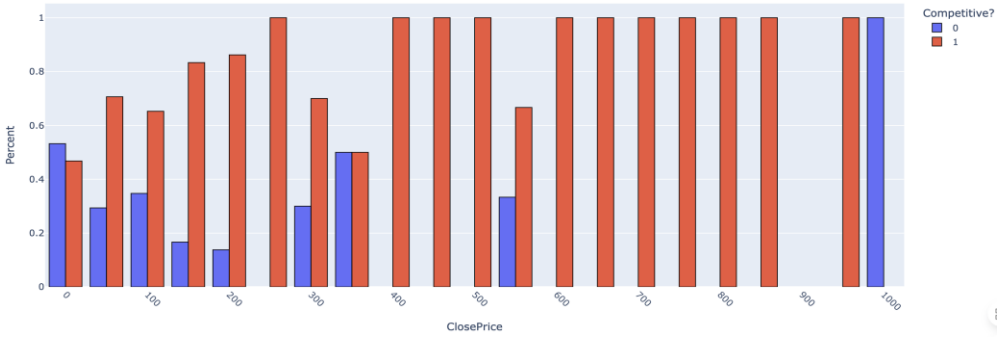


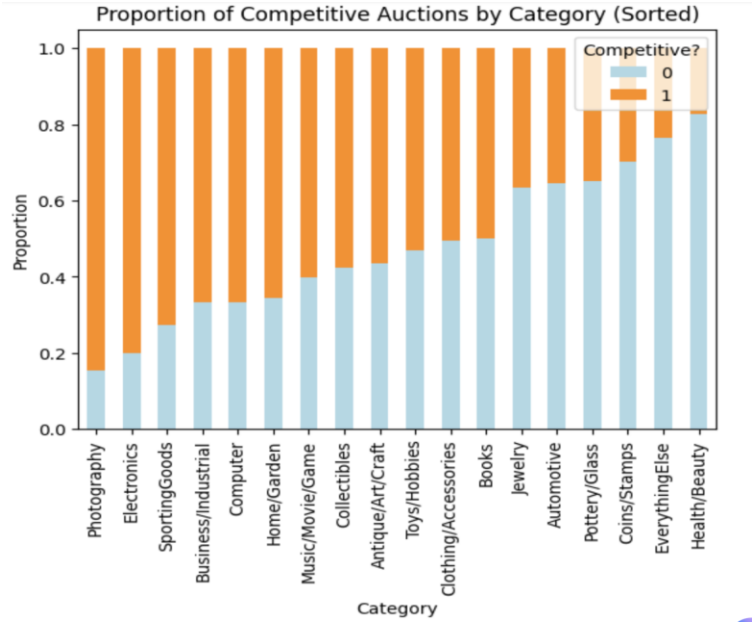


Bivariate Analysis (vs. Target: Competitive?)

- Competitive auctions tend to occur at very low or very high *sellerRatings*, suggesting a non-linear relationship. They are also more common for 5- and 10-day durations and when *OpenPrice* is below $100. While low *ClosePrices* can occur in both outcomes, higher closing prices are almost always associated with competitive auctions (see top 2 graphs in next page)

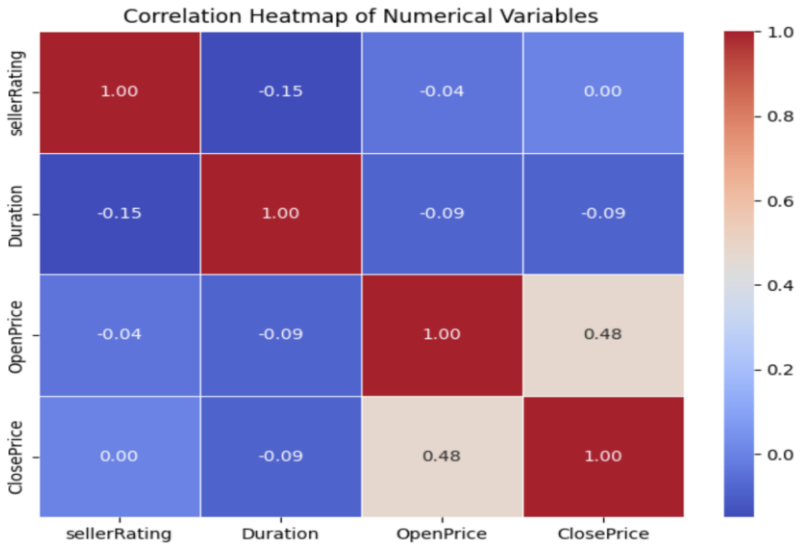






- Certain categories—*Photography, Electronics, SportingGoods, Business/Industrial,* and *Computer*—show the highest competitiveness, while *Collectibles, Music/Movie/Game, Antique/Art/Craft,* and *Toys/Hobbies* are balanced, and other categories are mostly non-competitive (see: graoh above).
- Auctions in GBP are more competitive, EUR auctions are balanced, and US auctions are least competitive. Competitiveness is highest for auctions ending on Monday and Thursday, lower on weekends (especially Saturday), with midweek days showing a balanced mix.

Multivariate Analysis



- There is a **strong positive correlation** between OpenPrice and ClosePrice, indicating potential **multicollinearity** issues in modeling.

K-Nearest Neighbors (KNN)

Approach

We trained and evaluated three KNN models:

1. A **baseline model** with default parameters (K=5)
2. A **tuned model using all features** (to assess full predictive potential)
3. A **tuned model excluding ClosePrice** (to simulate live prediction scenarios where closing price is not yet known)

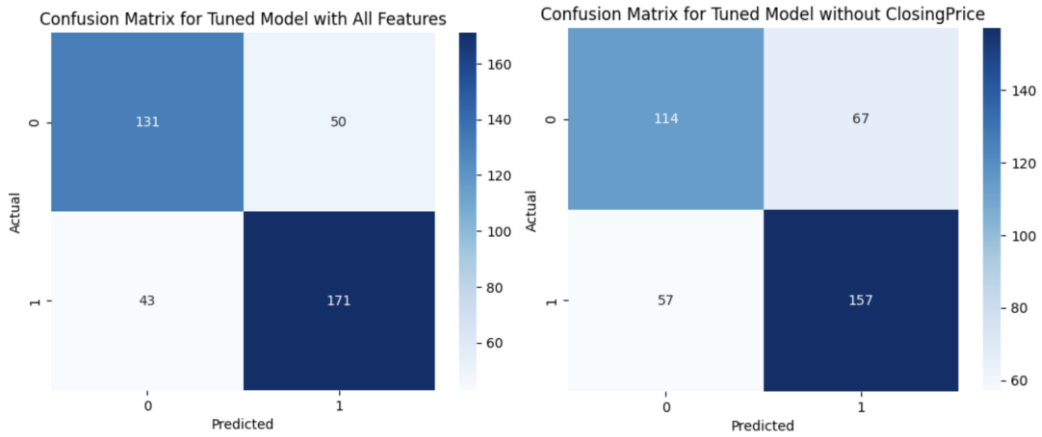
Hyperparameter tuning with cross-validation was applied to optimize the number of neighbors, distance metric, and weighting scheme, aiming to maximize validation performance while controlling overfitting.

Key Findings

Below is a summary of each model's performance, including accuracies, precisions, recalls, etc.

Model	Train	CV	Test
Baseline (All Features Default Params)	82.24%	-	73.92%
Tuned (All Features, Hyperparameter CV)	99.37%	77.55%	76.46%
Tuned (All Features except ClosePrice, GridSearch CV)	91.82%	68.86%	68.61%

Model	Precision	Recall	F1-score
Baseline (All Features)	0.74	0.74	0.74
Tuned (All Features, Hyperparameter CV)	0.69	0.69	0.69
Tuned (All Features except ClosePrice, GridSearch CV)	0.69	0.69	0.69



Model Selection Scenarios

In practical prediction scenarios, the availability of the ClosePrice variable depends on the timing of the prediction. When the goal is to forecast future competitiveness, for example, predicting whether the market will be competitive tomorrow or next week. In that situation, the future closing price is not yet known. In such cases, we must rely on the Tuned model without ClosePrice, which is designed to perform reliably even without this variable. Conversely, when conducting historical analysis or post-market evaluation, the ClosePrice is already available, allowing us to use the Tuned model with all predictors to achieve higher predictive accuracy. This approach ensures that our model selection aligns with real-world data availability, maintaining both practical applicability and analytical rigor.

Discussion / Answers to Questions

Predicting competitive auctions before the auction closes is feasible but less accurate, achieving around 69% test accuracy when ClosePrice is unavailable. The Closing Price proves to be a strong predictor of auction competitiveness and omitting it reflects the real-world limitation faced in live prediction settings. Nevertheless, the model remains valuable for guiding sellers and auction managers by identifying listings that are likely to attract competition, enabling proactive adjustments to starting prices, auction duration, or promotional strategies ahead of auction close. To further enhance predictive performance, it is essential to tune K and other hyperparameters to balance bias and variance, particularly given the high-dimensional nature of the dataset. This ensures the model remains both practical and generalizable across different auction scenarios.

Decision Trees

Approach

Decision Tree models were trained under three configurations: a baseline model using all features to assess overall predictive power and feature importance; a baseline variant excluding *ClosePrice* to reflect real-world prediction where final prices are unknown; and a tuned model using only the top predictors (*OpenPrice* and *SellerRating*), optimized through GridSearchCV with cross-validation. Each tree used a minimum of 50 samples per leaf to control overfitting and balance model complexity with accuracy.

Key Findings

Below is a summary of each model's performance, including accuracies, precisions, recalls, etc.

Model	Train	CV	Test Accuracy
Baseline Exploratory Model (All Features)	83.51%	-	80.76%
Baseline Predictive Model (All Features except ClosePrice)	73.80%	-	69.60%
Tuned Model	74.06%	72.92%	72.15%

Model	Precision	Recall	F1-score
Baseline Exploratory Model (All Features)	0.81	0.81	0.81
Baseline Predictive Model (All Features except ClosePrice)	0.70	0.70	0.70
Tuned Model	0.72	0.72	0.72

Discussion / Answers to Questions

Baseline Exploratory Model (All Features)

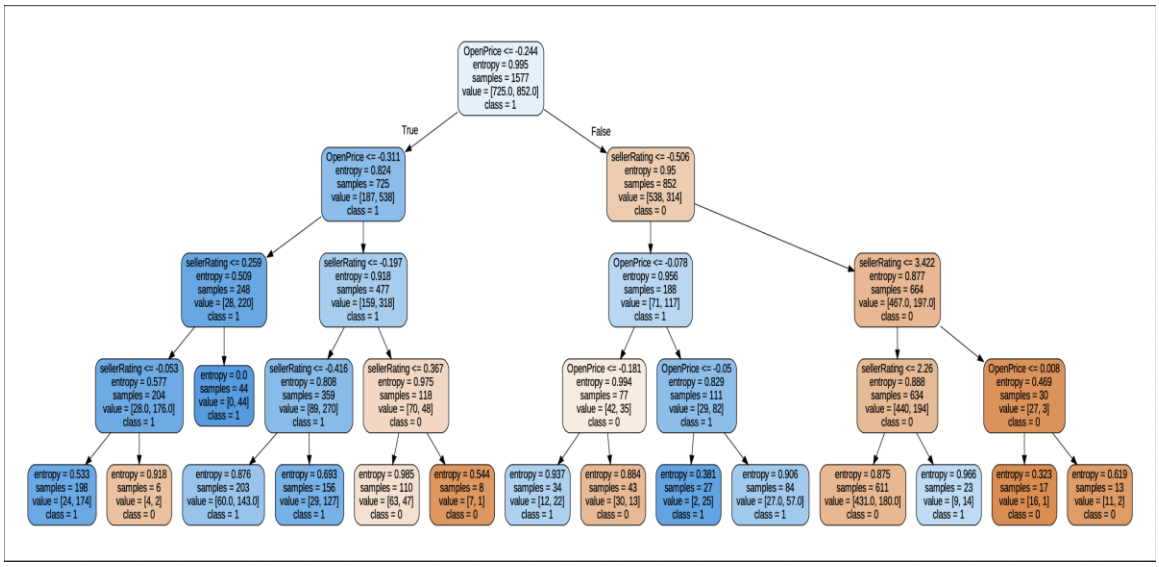
The first model, the Baseline Exploratory Model, used all available features. While it performed well in terms of accuracy, ClosePrice cannot be used for predicting new auctions because it is only observed after the auction ends. Including it would constitute "cheating" for future predictions. Therefore, the model was retrained without ClosePrice, forming the Baseline Predictive Model.

Baseline Predictive Model (All Features except ClosePrice)

Observations show that OpenPrice remains the dominant predictor (importance 0.58), with lower starting prices increasing the likelihood of competitive auctions. SellerRating has a moderate effect (0.31), indicating that more reputable sellers slightly improve competitiveness. Duration has minor influence (0.07), while Category (Toys/Hobbies) and Currency (EUR) have minimal impact (<0.03). Other categories, currencies, and endDay variables contribute nothing to predictive power. This suggests that for practical prediction, only a few predictors—OpenPrice, SellerRating, and optionally Duration—are needed.

Tuned Model

Based on the feature importances from the predictive baseline model, the top predictors are OpenPrice (0.582), SellerRating (0.309), and Duration (0.074). For simplicity and interpretability, only OpenPrice and SellerRating are retained in the tuned model. This ensures that predictions for new auctions rely on the most informative and actionable features, while reducing complexity and avoiding reliance on variables with minimal impact.

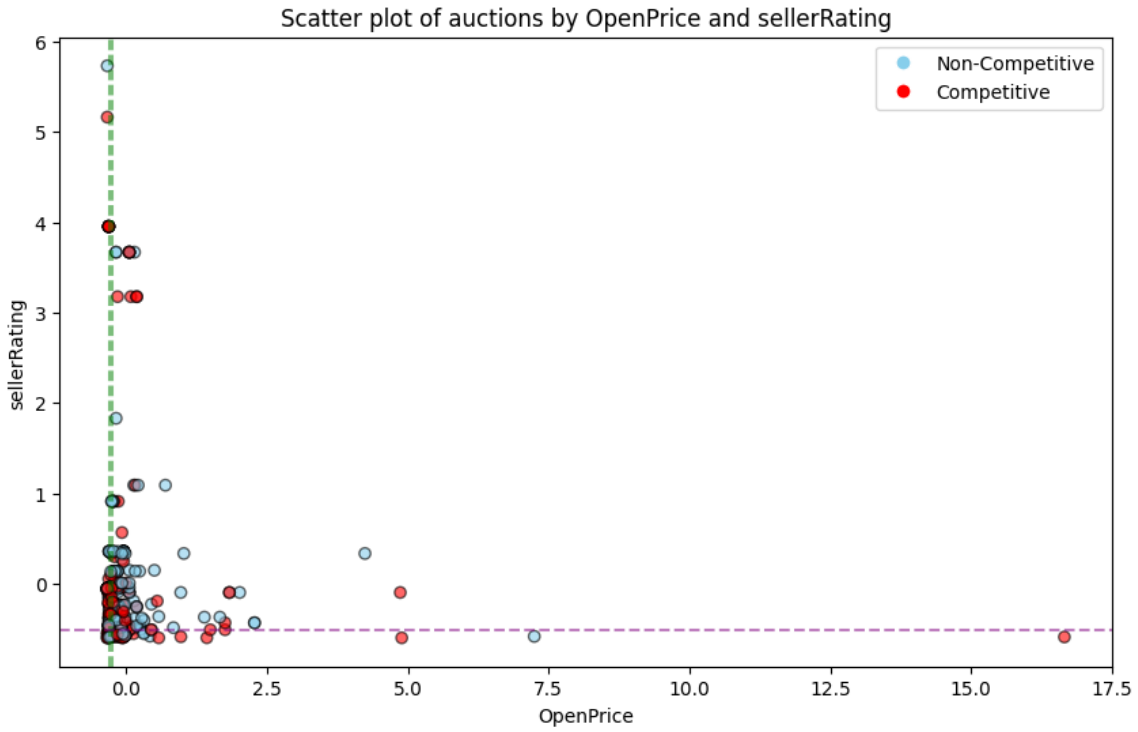


The full class tree of Tuned Model

- **Key Findings from the Scatter Plot**

  OpenPrice (importance 0.670) is the dominant predictor. As displayed in the figure below, the first split, represented by the green vertical line, separates many non-competitive auctions, though some competitive auctions remain on both sides. SellerRating (importance 0.339) provides secondary separation; the second split, shown by the purple horizontal line, further

isolates non-competitive auctions in the top-right quadrant, but some competitive auctions still remain mixed.



- **Implications**
  OpenPrice remains the primary driver of auction competitiveness, with seller reputation providing moderate additional separation. Applying GridSearchCV to tune the tree using only the top predictors (OpenPrice and SellerRating) improved test accuracy from 69.6% (Baseline Predictive Model) to 72.2%, demonstrating better generalization while keeping the model simple and applicable to new auctions where ClosePrice or endDay are unknown.

Recommendations for Sellers

1. Set a lower opening price to maximize the chance of multiple bids.
2. Consider seller reputation as a secondary factor.
3. Focus less on Duration, Category, Currency, or endDay, as these have minimal impact on auction competitiveness.

Comparative Analysis: KNN vs Decision Tree

• Predictive Performance:

- Decision Trees consistently outperform KNN, achieving ~81% accuracy with all features and ~72% for live predictions using top features, whereas KNN peaks at ~77% with all features and drops to ~68% without *ClosePrice*.

Interpretability:

- Decision Trees are more interpretable: feature importance and decision paths are transparent, making it easier to communicate insights to stakeholders.
- KNN is less interpretable, as predictions depend on distances to neighboring points without clear rules.

Trade-offs:

- KNN is sensitive to feature scaling and high dimensionality, and its performance drops when key features are unavailable.
- Decision Trees strike a strong balance between **accuracy, interpretability, and robustness**, providing reliable predictions and actionable insights for sellers.
- Overall, Decision Trees offer higher business value by combining predictive power with explainable results, making them the preferred approach for this application.

Generative AI Usage and Team Contributions

Gen AI Usage

- Used **ChatGPT (GPT-5)** as a learning and writing support tool to improve report clarity, structure, and formatting.
- Helped refine **EDA summaries** and improve chart readability (e.g., axis labels, layout).
- Clarified how to use **Pipeline with GridSearchCV** in KNN to prevent data leakage and ensure reproducible workflows.
- Assisted in explaining **bias–variance trade-offs** and model behavior.
- All **coding, analysis, model design, and interpretation** were independently performed by the project team.
- GenAI outputs were **reviewed, edited, and validated** before inclusion.

Team Contributions

- Zimeng – EDA
- Peter – Data preprocessing & helped in modelling sections
- Zoe & Yaojie – KNN
- Harrison – Decision Trees
- Laana – Solution Approach, helped across all sections