E-bay Auction Competitiveness Project Report

Part 1: Data Preprocessing and Exploratory

1.1 Data Preprocessing

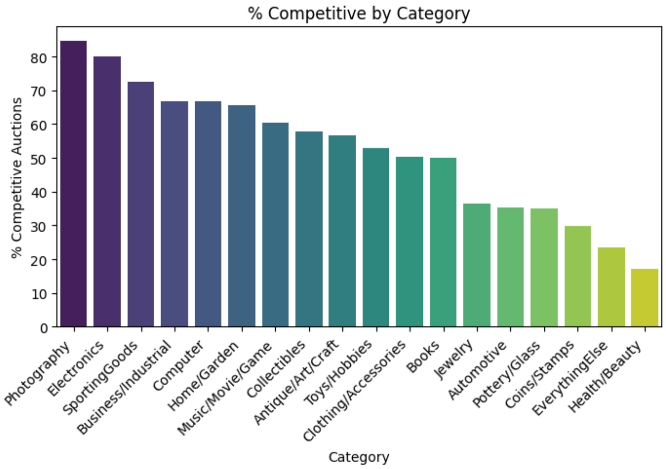
1.1.1 Data Cleaning & Preparation

The dataset contains 1,972 auctions, with no missing values in any field. All features are complete, so no imputation was required.

- **Features Identified:**
  - **Numerical:** SellerRating, Duration (auction length in days), OpenPrice, ClosePrice.
  - **Categorical:** Category (product type), Currency, EndDay (day of week auction closed).
  - **Target:** Competitive (1 if ≥2 bids, else 0). Originally stored as numeric (0/1), converted to Boolean for clarity.
- **Preprocessing Steps:**
  - One-hot encoding applied to all categorical variables, keeping all levels (no baseline dropped) to preserve interpretability in tree/k-NN models.
  - Continuous features were examined for extreme values. Instead of removal, outlier flags were created to preserve data for EDA while allowing clean subsets for model training.
  - **Transformations:**
    - **Log scaling (log1p)** applied to highly skewed features (OpenPrice, ClosePrice, SellerRating) to normalize distributions.
    - **Price ratio** (ClosePrice / OpenPrice) was calculated to capture relative bidding dynamics rather than absolute prices. A log transform further helped us fit it into a box plot.
    - **Seller tiers** created by quantile binning of SellerRating into four groups (Low, Mid, High, Top).

1.2 Exploratory Data analysis

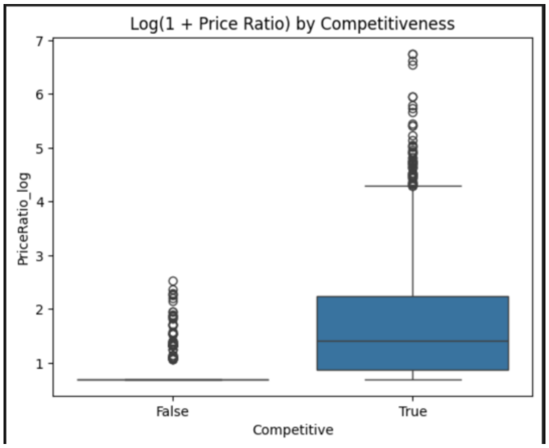
- **Category:**
  - Competitiveness is not uniform across item categories.
  - High-value categories such as Jewelry, Electronics, and Collectibles show much higher competitiveness rates, while Books and Media categories often attract only a single bid.
  - This suggests bidders are more likely to engage in competitive bidding when items are perceived as rare, premium, or high-demand.



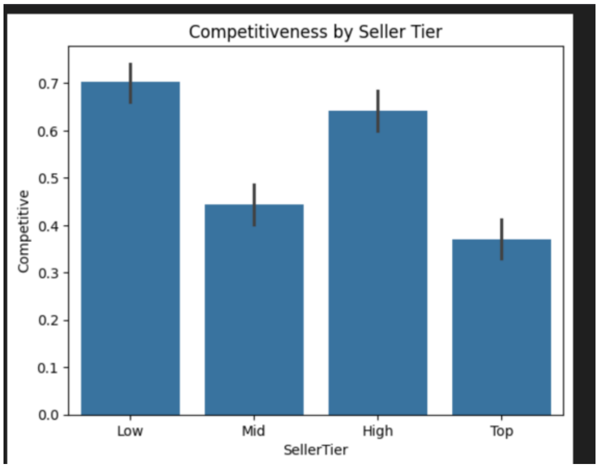
- **Price Ratio (Closing vs. Opening Price):**

Defined as price ratio = (1+ClosePrice)/(1+OpenPrice). If the ratio close to 1, meaning the closing price close to opening price. If price ratio is much larger than 1, meaning the price grew a lot, suggesting a competitive auction.

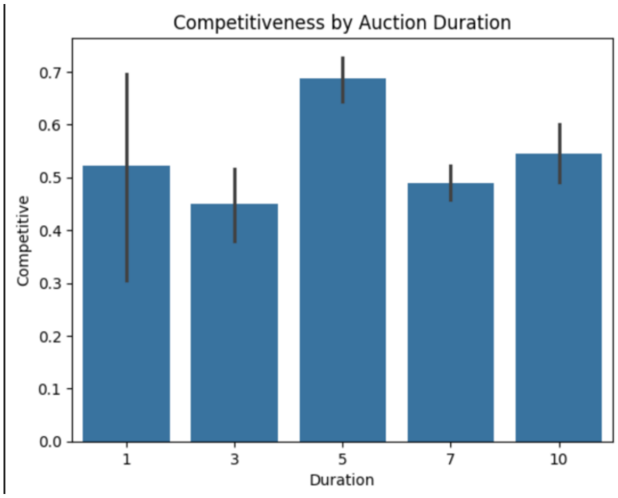
  - Auctions where the final price was much higher than the opening price were consistently competitive.
  - Non-competitive auctions often had a closing price very close to the opening price, indicating limited or no bidding activity.
  - This confirms that competitiveness is better explained by relative price growth than by absolute price alone.



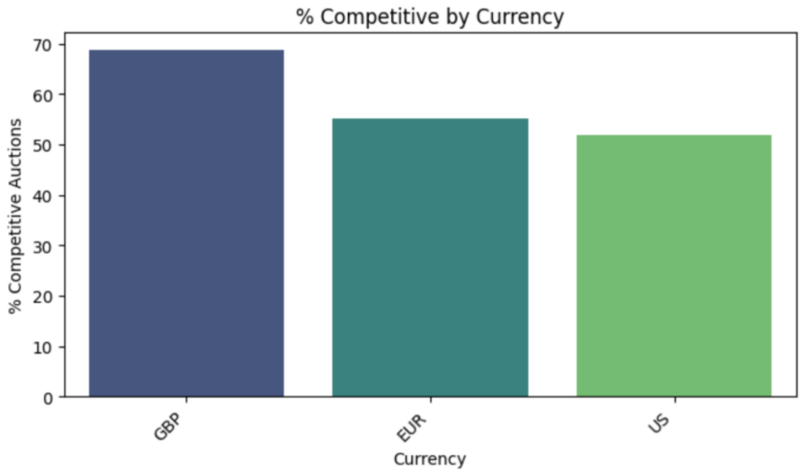
- **Seller Rating & Tiers:**
  - Sellers with higher reputation (Top tier) ran a significantly larger share of competitive auctions compared to low-rated or new sellers.
  - This highlights the trust factor: buyers are more willing to bid actively when the seller is perceived as credible.



- **Auction Duration:**
  - Auctions with longer durations (7–10 days) show higher competitiveness compared to short auctions (3 days).
  - This supports the idea that time exposure increases bidder participation, giving more potential buyers a chance to notice and bid.



- **Currency & End Day:**
  - Initial analysis indicates minimal standalone impact on competitiveness.
  - However, interaction effects may exist (e.g., certain categories closing on weekends might behave differently). These merit deeper modeling rather than visual EDA alone.

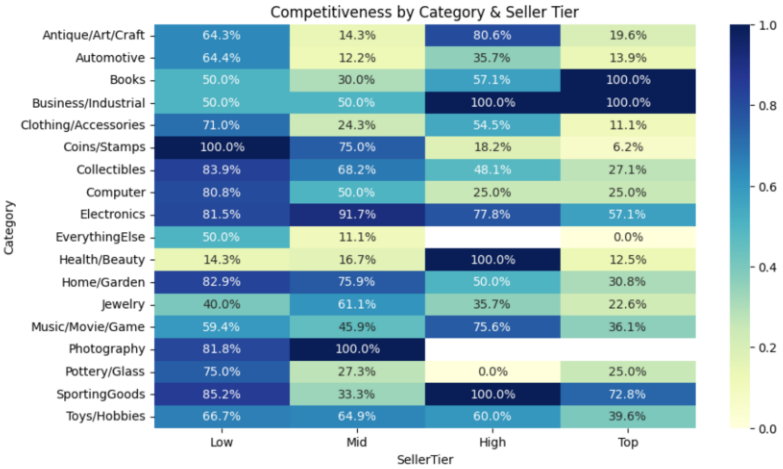


Visual Insights

Bar charts: Proportion of competitive auctions across categories, seller tiers, and auction durations.

Boxplots: Distribution of log-transformed prices and price ratio by competitiveness, showing clear separation in medians.

Heatmaps: Competitiveness by Category × Seller Tier, which revealed that high-reputation sellers dominate competitive categories like Electronics and Jewelry.



1.3 Key Findings & Takeaways

1. **Relative growth (price ratio)** is more predictive of competitiveness than raw prices, a small opening bid that escalates is a hallmark of competitive bidding.

2. **Seller reputation matters significantly**: established, top-rated sellers enjoy higher bidder trust, leading to more active auctions.

3. **Auction duration is a controllable factor**: sellers who list longer auctions increase chances of competition, likely due to greater exposure.

4. **Category effects are strong**: certain item types inherently attract competitive behavior, possibly due to scarcity or collectability.

5. **Interaction effects matter**: competitive patterns emerge when combining seller reputation, item category, and auction length, not just in isolation.

Part 2: Prediction Models

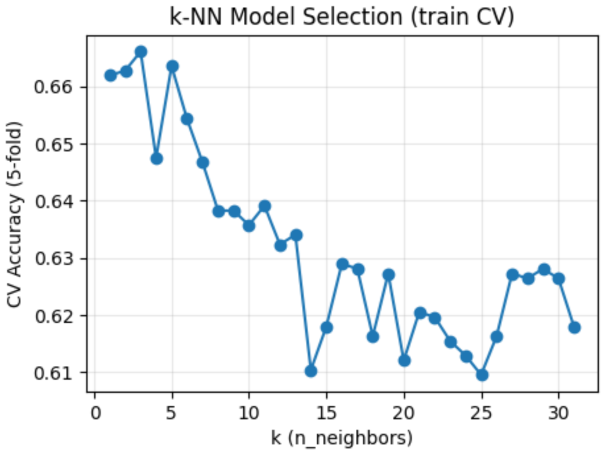
2.1 K-Nearest Neighbours (k-NN)

Selecting competitiveness as the target variable:

- **Objective**: Predict whether an auction will be Competitive (≥ 2 bids) vs Not Competitive (< 2 bids) at, or before, listing time.
- **Target variable**: competitive (binary).
- **Feature policy:** Only pre-listing/listing-time features were used (no post-auction leakage). Currencies were normalized into USD, categorical fields were converted into binary indicators, and numeric fields were standardized.
    - **Categorical**: Category, Currency, endDay, seller_tier
    - **Numeric**: sellerRating, Duration

Tuning k value:

- Conducted a 5-fold cross-validation on the training split only
- After tuning k, we found that the k value that peaks performance and the balance variance tradeoff is **k=5.**
- The curve shows the usual bias–variance pattern: very small k overfits (high variance), larger k smooths decision boundaries (higher bias).
- The train CV accuracy for the 5-fold is 0.687.



Hold out performance test (60%/40% split):

Metric	**Value**
Accuracy	0.674
Precision (Competitive ≥2 bids)	0.717
Recall (Competitive ≥2 bids)	0.777
F1 Score (Competitive ≥2 bids)	0.745

- The accuracy for the hold out performance test 54.1% of listings are Competitive (≥2 bids). A naïve majority baseline would be around 0.54 accuracy, our k-NN at around 0.67 exceeds that, so the model adds signal.

Confusion Matrix:

	Predicted: Not Competitive	**Predicted: Competitive**
True: Not Competitive (< 2 bids)	128	123
True: Competitive (≥2 bids)	89	311

Interpretations:

The k-Nearest Neighbors model (k = 5) demonstrates reasonable predictive strength, correctly identifying roughly two-thirds of listings overall. Its recall of 0.777 indicates that it successfully captures nearly 78% of truly competitive auctions, while its precision of 0.717 shows that about 72% of listings flagged as competitive are.

However, there remains a notable false-positive rate, as 123 non-competitive listings were incorrectly classified as competitive. If downstream promotional actions are costly (e.g., homepage placement, discounting), this misclassification rate could be expensive. Conversely, if the business impact of missing a competitive auction is greater (lost engagement or take-rate), the current configuration prioritizing recall is acceptable.

The model is good at finding likely competitive auctions, but it will also flag some non-competitive ones, this is a tolerable trade-off if the upside from capturing competitive auctions outweighs promo waste.

Business Implications:

1. **Operational Use**
   - The k-NN model can act as a pre-screening tool to prioritize listings for promotional placement or early visibility boosts.
   - It is most effective when the goal is to capture as many competitive auctions as possible, rather than to perfectly avoid non-competitive ones.
2. **Strategic Thresholding**

   - Because recall is higher than precision, the default threshold favors inclusivity.
   - If marketing or promotion costs are significant, increase the classification threshold (favoring precision).
   - If missed competitive auctions carry greater opportunity cost, lower the threshold (favoring recall).

3. **Interpretability and Maintenance**
   - k-NN is non-parametric and intuitive: similar listings yield similar competitiveness predictions, which aligns with how sellers and analysts reason about comparable items.
   - However, the model is computationally heavier and sensitive to scaling; it should be monitored as new categories or currencies are introduced.

4. **Limitations**
   - k-NN does not provide feature importance scores, so business interpretation of drivers requires auxiliary analysis.
   - The model's moderate accuracy (around 0.67) implies it should not be the sole decision-maker. Instead, integrate it with business rules or subsequent models.

5. **Actionable Recommendations**
   - Pilot deployment: use predictions to inform where to allocate promotional inventory or which listings to highlight.
   - Cost–benefit calibration: link model thresholds to the monetary value of correctly versus incorrectly classifying a listing.
   - Continuous learning: periodically retrain as new data accumulate, ensuring local distance patterns remain valid.

The k-NN model demonstrates meaningful predictive value and captures the behavioral clustering of competitive auctions. It provides a practical, data-driven foundation for targeting, but its outputs should be paired with cost-based thresholds and complemented by interpretable follow-up models for sustained business use.

2.2 Decision Tree

After the K-nearest-neighbor model analysis, a decision tree classifier was implemented to predict auction competitiveness using all available predictors. The tree's structure revealed intuitive and interpretable patterns consistent with the EDA findings.

Feature Importance:

Feature importance analysis revealed that OpenPrice and ClosePrice together accounted for over 95% of the model's total importance, while SellerRating contributed approximately 5%. Other predictors, such as Category, EndDay, and Duration, exhibited minimal influence. This pattern persisted after cross-validation when tuning the model's parameters for new auctions (excluding ClosePrice as a predictor). In this case, OpenPrice and SellerRating jointly explained over 95% of the total importance, with OpenPrice's contributing 65% and

SellerRating 32%. These findings highlight the model's strong preference for continuous variables with substantial predictive power.

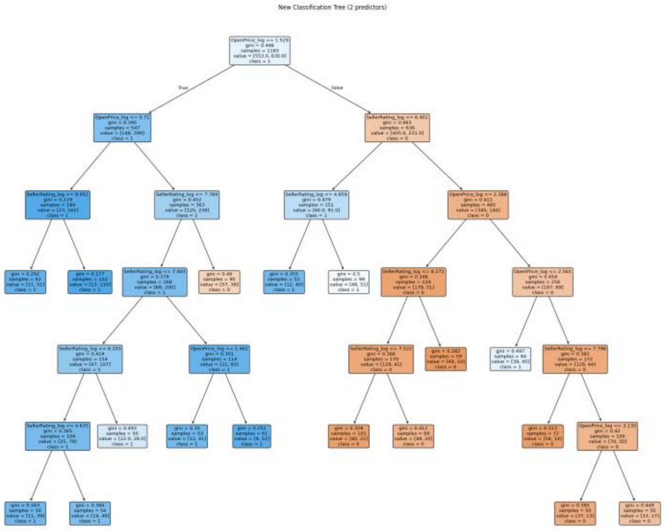
Key Findings:

- Obvious Findings:
- The opening price and closing prices pose the dominant impact on an auction being competitive or not.
- Auctions of low prices are mostly competitive as they attract more bidders into the auction.
- Auctions which saw large differences between opening and closing prices are more likely competitive as more bids are happening bidding up the price.
- Unexpected Findings:
- Despite the expectation that more competitive auctions might last longer, and that during weekends people shall have more time to focus on bids and attend auctions, neither Duration nor EndDay appear in the decision rule of the tree. Similarly, despite the expectation that certain types of items should be more attractive to bidders, Category does not appear in the decision rule.
- Auctions starting at high opening prices are more likely to be competitive when the sellers have smaller past-auction scores. This is not expected as more experienced sellers are expected to catch more attention from bidders, and be more successful in increasing the competitiveness of auctions. However, it might be that the low-rated sellers use aggressive pricing to compete, and the resulting lower price attracts more bidders to enter the auction.

Model Interpretation:

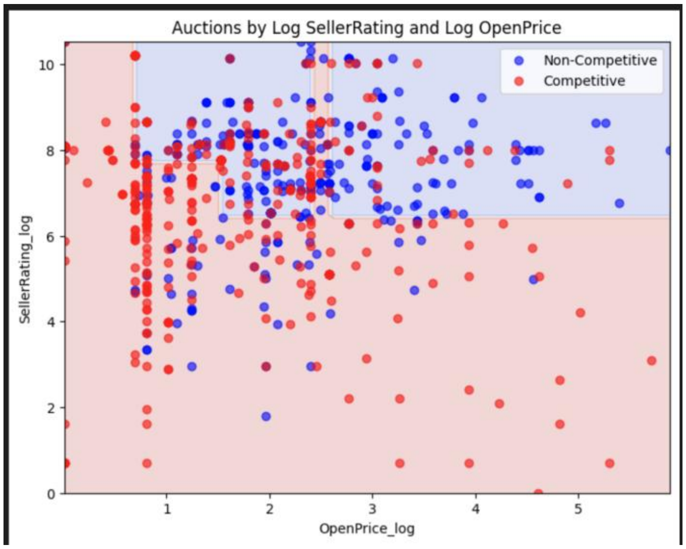
To ensure interpretability, a simplified tree was refitted using only OpenPrice and SellerRating—the two most influential features for the decision boundary visualization. The resulting splits were logical:

- Auctions with lower starting prices or lower-rated sellers tended to be more competitive, likely because sellers with lower ratings use aggressive pricing to attract bidders.
- Auctions with high starting prices and high-rated sellers were more likely to be non-competitive.
- The interactive impact of competitive pricing and a low seller rating together increases the potential of an auction being competitive.


New Classification Tree (2 predictors)
OpenPrice_log <= 1.529 gini = 0.498 samples = 1183 value = [553.0, 630.0] class = 1
True
False
OpenPrice_log <= 0.71 gini = 0.396 samples = 547 value = [148, 399] class = 1
SellerRating_log <= 6.401 gini = 0.463 samples = 636 value = [405.0, 231.0] class = 0
SellerRating_log <= 8.052 gini = 0.219 samples = 184 value = [23, 161] class = 1
SellerRating_log <= 7.769 gini = 0.452 samples = 363 value = [125, 238] class = 1
SellerRating_log <= 4.859 gini = 0.479 samples = 151 value = [60.0, 91.0] class = 1
OpenPrice_log <= 2.388 gini = 0.411 samples = 485 value = [345, 140] class = 0
gini = 0.292 samples = 62 value = [11, 51] class = 1
gini = 0.177 samples = 122 value = [12, 110] class = 1
SellerRating_log <= 7.003 gini = 0.379 samples = 268 value = [68, 200] class = 1
gini = 0.48 samples = 95 value = [57, 38] class = 0
gini = 0.355 samples = 52 value = [12, 40] class = 1
gini = 0.5 samples = 99 value = [48, 51] class = 1
SellerRating_log <= 8.271 gini = 0.346 samples = 229 value = [178, 51] class = 0
OpenPrice_log <= 2.565 gini = 0.454 samples = 256 value = [167, 89] class = 0
SellerRating_log <= 6.203 gini = 0.424 samples = 154 value = [47, 107] class = 1
OpenPrice_log <= 1.462 gini = 0.301 samples = 114 value = [21, 93] class = 1
SellerRating_log <= 7.522 gini = 0.366 samples = 170 value = [129, 41] class = 0
gini = 0.282 samples = 59 value = [49, 10] class = 0
gini = 0.497 samples = 84 value = [39, 45] class = 1
SellerRating_log <= 7.796 gini = 0.381 samples = 172 value = [128, 44] class = 0
SellerRating_log <= 4.635 gini = 0.365 samples = 104 value = [25, 79] class = 1
gini = 0.493 samples = 50 value = [22.0, 28.0] class = 1
gini = 0.35 samples = 53 value = [12, 41] class = 1
gini = 0.252 samples = 61 value = [9, 52] class = 1
gini = 0.329 samples = 101 value = [80, 21] class = 0
gini = 0.412 samples = 69 value = [49, 20] class = 0
gini = 0.313 samples = 72 value = [58, 14] class = 0
OpenPrice_log <= 3.135 gini = 0.42 samples = 100 value = [70, 30] class = 0
gini = 0.343 samples = 50 value = [11, 39] class = 1
gini = 0.384 samples = 54 value = [14, 40] class = 1
gini = 0.385 samples = 50 value = [37, 13] class = 0
gini = 0.449 samples = 50 value = [33, 17] class = 0

Visualization and Transformation:

Decision boundaries were plotted over a scatter plot of the two predictors. Without transformation, the data were tightly clustered due to outliers, so log-transformations were applied to both continuous predictors to improve clarity and alignment with the EDA process. The visualization confirmed that the model performed well in separating extreme regions of opening price and seller rating (clearly competitive vs. non-competitive auctions), though overlap remained in mid-range values where other unmodeled factors likely play a role. With a two-way scatter plot with two predictors, there is not enough dimension to visualize the influence of other predictors.



Model Evaluation:

The best tree with cross-validated parameter-searching on all predictors except closing price achieved 71.5% accuracy on the test set, only slightly lower than the 72.7% on the training set, suggesting minimal overfitting. Among actual competitive auctions, 77.6% were correctly predicted, and among actual non-competitive auctions, 62.9% were correctly predicted. Among the predicted competitive auctions, 72.3% are actually competitive, and among the predicted non-competitive auctions, 70.3% are actually non-competitive.

Confusion Matrix:

	Predicted: Not Competitive	**Predicted: Competitive**
True: Not Competitive (< 2 bids)	222	131
True: Competitive (≥2 bids)	94	342

Insights and Recommendations

The analysis suggests that OpenPrice and SellerRating alone are sufficient for reasonably accurate prediction of auction competitiveness. Additional features provided limited incremental predictive power, consistent with decision trees' inherent feature selection capability: it prioritizes the most informative predictors to determine the splits, and assigns greater importance to continuous variables compared with dummy variables with only two levels.

Recommendations for sellers based on the insights are:

- Set an optimal opening price
- The decision tree shows that extremely high or low opening prices strongly influence competitiveness. Medium-range prices tend to be competitive because they attract more bidders without scaring them away.
- Sellers should set the opening price in a range that balances attractiveness and value—not too high to discourage bidders, not too low to signal low value.
- Sellers could analyze past auctions for similar items to find the suitable start price.
- Maintain or improve seller rating
- Higher seller ratings correlate with more competitive auctions.
- Buyers tend to trust sellers with good reputations, leading to more bids.
- Recommendation: Focus on good service, accurate descriptions, and timely shipping to maintain a strong rating.
- Choose time of the auction based on convenience or cost
- Variables like EndDay, Duration, and Currency had little effect on competitiveness in the tree.

Part 3: Comparative Analysis

The following comparative analysis evaluate the interpretability and predictivity of the KNN model and the decision tree model, and finally, we compare their business applications. Both models are trained and tested on the same preprocessed dataset for consistency

3.1 Interpretability

KNN

The KNN model requires little assumption, but at the same time, provide limited interpretability. The model's accuracy, recalls and precision on the test dataset identify competitiveness reasonably effectively, making it suitable for prediction but not inference, i.e, finding the causes of competitiveness.

Decision Tree

The decision tree provides clear insights. Under the best tree, we can see that Low opening prices and high seller ratings increased competitiveness, while extremely high prices discouraged bidding activity. The limited depth of the model minimize overfitting and also provides easier interpretation due to its simplicity. For example, Decision such as SellerRating <= 601.5, and SellerRating <=128 results in competitive, gives clear actionable insights to sellers.

In summary, the decision tree model has much better interpretability than KNN.

3.2 Predictivity

Model	**Test Accuracy**	**Precision (Competitive)**	**Recall (Competitive)**	**F1 Score**	**Key Predictors**
k-NN (k=5)	**0.674**	0.717	**0.777**	0.745	Category, SellerRating, Duration, OpenPrice
Decision Tree	**0.715**	0.723	0.776	**0.749**	OpenPrice, SellerRating,

The test accuracy is the most important indicator of predictivity. The decision tree outperformed KNN on the test set with a test accuracy of 0.715 verses 0.674. The precision and the recall of the the two models are very similar, it indicates that both models were able to categorize a competitive auction reasonably well.

Overall, the decision tree model has much better predictivity than KNN since it has a higher test accuracy.

3.3 Trade off between predictivity and interpretability

Criterion	Decision Tree	KNN
Test Accuracy	0.674	0.777
Interpretability	Low	High
Computation	Costly for large data set	Efficient and Scalable

Overall, decision tree is superior in both predictive power and interpretability. Moreover, decision tree is les computational heavy.

3.4 Business Usage

Decision Tree's explicit rules making it ideal for decision making. With better test accuracy, and similar precision and recalls, the decision tree model has better business value than KNN.

Business usage on the seller level

- KNN
  - provide sellers real time predictive competitiveness before listing
- Decision Tree
  - provide not only prediction, but also insights into thresholds on seller ratings and opening price
- Combine Usage
  - KNN could be used for identifying competitiveness, then decision tree can provide insight into why these listings are competitive, and guides sellers to future decisions