Alternative Assets

knawa, dgarci10, rjawle, bhurd

Goal

The realm of Alternative Assets (AA) is notorious for high barriers of entry. The popularity of sports memorabilia, specifically baseball, basketball, and football cards, has been increasing over the past few years. As this market grows, it has become increasingly profitable to trade these assets. This is why we created a predictive model for each type of sports card that looks at the AA's previous selling points and other factors to determine the best current price. With our model's estimated current price, we have the knowledge to evaluate whether or not the sports memorabilia is overvalued or undervalued and buy or sell based on when we deem most fit.

Data

Our data came from the Professional Sports Authenticator (PSA) website and we retrieved information on baseball, football, and basketball cards sold. By combining that with statistics of the athlete (which we got from sports-reference.com), we had a dataset that contained the history of prices at which a card was sold and significance of the player. We had to remove items that were only sold once since we were not able to create a proper history of prices. In addition, cards that represented more than one player also had to be removed since we did not have an accurate way of combining the players' statistics into one metric. We had an average of 5855 data points for each sport which were used for the 3 different models.

Model+Evaluation Setup

We used multiple regression as our prediction model, since our ultimate goal was to predict the price of a card if it were to be sold now given a number of different variables that the price was dependent on. The 15 independent variables are listed below:

0: Year card was created
1: Current grade of the card (quality of the card from 0.0 - 10.0)
2: Is Hall of Fame (0 - was not Hall of Fame, 1 - was Hall of Fame)
3: Average Price of the card
4: Percent Price Change per Year of the card
5: Highest Price at which it was sold
6: Lowest Price at which it was sold
7: Count of number of times sold
8: The number of awards the player on the card won
9: The Oldest Price of the card
10: Oldest Trade Date of the card
11: Most Recent Trade Date of the card
12: Important Player Stat #1 (dependent on the sport)

13: Important Player Stat #2 (dependent on the sport)
14: Important Player Stat #3 (dependent on the sport)

Since the three sports that we looked at for predicting card prices have wildly different key statistics, we decided to collect different statistics for each sport (Variables 12-14). For baseball, we collected Wins over Replacement (WAR), Homeruns (HR), and Runs Batted In (RBI). For basketball, we collected Total Games Played, Points per Game (PPG), and Assists per Game (APG). For football, since statistics are completely different depending on a player's position, we omitted the statistics section entirely.

We ran the regression separately for each sport because of the differences in the statistics variables for each sport. We determined training and testing data using an 80/20 train/test split. In addition, we evaluated our results for every possible combination of variables to see which were actually needed to optimize our prediction model.

Results and Analysis

Claim #1: The most correlated variables with respect to recent price are: average price, lowest price, highest price, and oldest price.
Support for Claim #1: In order to get an understanding into which variables to choose for multiple regression a correlation matrix of all the data points was used Individual correlation matrices for each sport were also created to see if any significant differences could be concluded. However, every sport followed the claim as seen below in the chart.

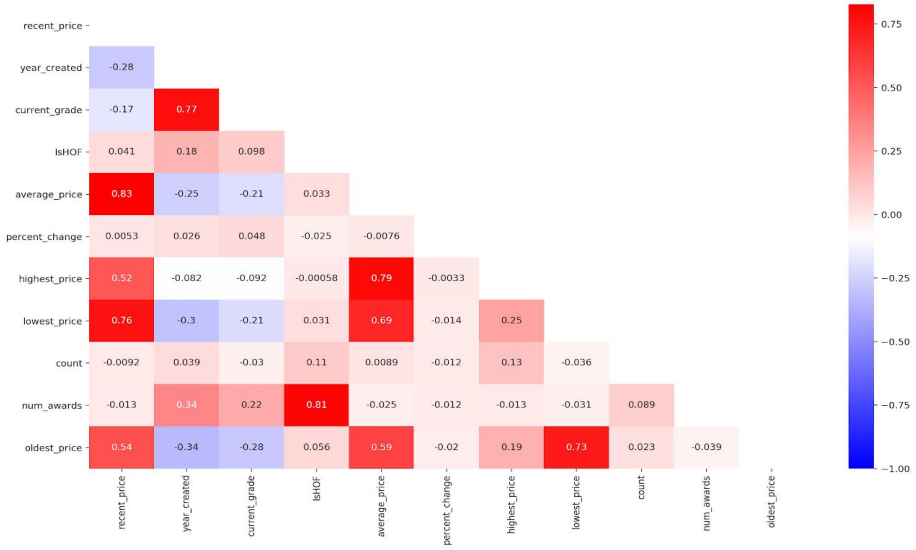

Figure 1: Correlation Matrix of All Data Points

Claim #2: By including only 9 of the 15 variables, we were able to optimize our regression model, signifying that those variables were the most important in predicting the price of the card.

Support for Claim #2: As previously stated, we ran the regression on all possible combinations of the 15 variables. When evaluating these results, we wanted to look at the combination that produced results that most efficiently increased our model's R^2 value and most efficiently decreased training and testing MSE and the discrepancy between the two. By doing so, we discovered that the nine variables (yearCreated, grade, average price, percent price change per year, count, oldest trade date, most recent trade date, stat1, stat2) was the combination that best optimized our regression model.

Multivariable Regression Results Summary

Variables	R^2	Test MSE	Train MSE
Top 9	**0.5867034709210321**	**2,930,908.429045171**	**2,441,127.635897760**
Top 4 Correlated	**0.5724917544208122**	**3,255,110.709511859**	**2,328,575.38211199**

Claim #3: Most predictions have a normalized error of between -5 and 5.
Support for Claim #3: As shown in this figure, most of the normalized errors for the testing data (this for the baseball dataset) are low. However, there are a few outliers where the prediction is far different from the actual current price.

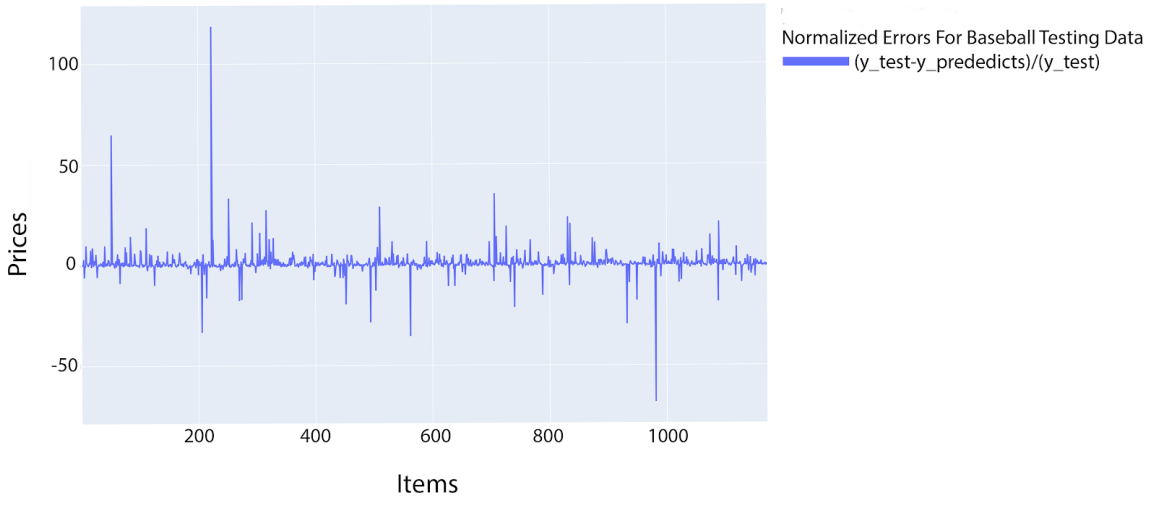

Figure 3: Normalized Errors