Baseball Project

By Charles Lascano, David Wiers, Jose Currea and Peyton Arana

Selected Topic

Analyzing the correlation between hitting performance and salary for MLB athletes using machine learning models and using this correlation to predict future data.

- Due to our shared interest in both baseball and finances, our project revolves around the topic of salaries in Major League Baseball(MLB) from 2009 to 2015 and the relationship that batting stats have in regards to a player's salary. We are going to look at the hitting stats for the years 2009-2014 to gain insights on both the hitting and salary stats. Once we have completed our extraction and transformation of the data, we will then test the 2015 data in a linear regression model to see how much of a correlation hitting performance has with the salary a player is paid. It is our hypothesis that there will be a very strong correlation between these two variables.

Reason for Selected Topic

- We chose this topic because we all shared an interest in both sports and finance and one question that utilized both of these interests was the salary of athletes.
- We then narrowed that initial interest down to the sport of Baseball and specifically hitters because unlike their pitching counterparts, hitters play far more frequently and are almost always the highest paid players on the team.
- In order to complete this project our first order of business was simply to look at our dataset, which gave us salaries from 1985-2015.
- To narrow our scope and for the sake of relevance, we decided to focus on a 5 year period of 2009-2014 as our training data and the singular year of 2015 as our testing data. This would allow us enough information we believe to analyze the data and test our hypothesis.

Data/Description of Source of Data

We pulled our Data from the Lahman's Baseball Datasets on SeanLahman.com. This website has baseball data from 1871-2021. This information had the hitting statistics we are using from 2009-2015 and the salary data from 2009-2015 as well. The hitting dataset had 22 columns and the salary dataset had 5 columns. They both shared two IDs which we could choose to merge on, and we chose to use the player_id column as they held unique IDs for each player in the dataset.

Questions

We have several questions that we are seeking answer through our data

- Based on the hitting performance of a player, is there a strong correlation between hitting performance and how much they are paid?
- Is it possible to predict the salary of a player using a linear regression model with the data given?
- Given a players salary for 2015, was this player overpaid or underpaid?
- What teams were the most successful at avoiding overpaying players, and who had the most players who were overpaid?

Description of Data Exploration

- The data exploration process included
    - Looking into what the dataset contained:
        - Reviewing the column headers using .head()
        - Looking to see that the timeline for the data matched with what we exported from Kaggle (2009-2014)
        - Looking at the number of records for each year using value.counts()
        - The number of unique values for each column using .nunique()
        - Using .describe to get the updated value of each column
        - Looking at the range of our target (Salary)
        - Reviewing max and mins of each column
        - Looked at the data-types that each column held in order to set up the machine learning model appropriately
        - Checked for null values
    - Reformatted the data and added missing columns
        - Columns added:
            - Singles, batting average, slugging %, OPS, total bases, BABIP, formulas included in README file
        - Bucketed the years (1-6)
        - Looking for outliers to remove or bin from the dataset
    - Created visualizations using seaborn and plt:
        - Heatmap to show correlation between columns
        - 2nd heatmap to show correlation of target (salary) to columns

Data Exploration/Analysis

- As far as Data Exploration/Analysis is concerned, our chosen dataset contains Baseball Statistics ranging from 1871-2015. However, as time has gone on, the sport of Baseball has continually progressed to include a wide range of Statistics. Consequently, when exploring our data, we have limited the range to 2009-2014, as this contains more relevant and pertinent statistics to record and make use of. In this Dataset, we are taking all recorded statistics of batting and salaries. With these two variables, we are using our machine learning model to analyze and give us a correlation. With this completed, we will use this machine learning model/correlation to predict future data in the year of 2015, which we will then compare to actual recorded information.

Description of the analysis phase of the project

The description of the analysis phase of our project is seen through the following link to our Machine Learning Model Information.md on our Repository

https://github.com/rclascano14/PROJECT/blob/main/ML_Information.md

Technologies, Languages, Tools and Algorithms used

- Jupyter Notebook
- Python
- Pandas
- Flask
- Psycopg2
- SQL
- PgAdmin4
- Tableau
- Project Presentation
- Project Dashboard
- Data: batting.csv, batting_filtered.csv, batting_salary.csv, salary.csv, salary_filtered.csv
- Sklearn - this package was imported to run the components inside of the machine learning models
- Logistic Regression - this was used for machine learning
- Linear Regression - this was used for machine learning

Machine Learning Model Descriptions

We preprocessed our preliminary data by placing it in the baseball_df dataframe. With our preliminary data formed into a dataset, we removed salary data from it and printed the shape of our dataset which resulted in X = (4793, 58) Y=(4793,). After, this we decided to split this preprocessed data into our training and testing datasets. We decided for our training dataset to be our baseball data from 2009-2014 and our testing data to be the singular year of 2015. This split was created through the following code, `X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=12, test_size=.20).` Lastly, the model choice we decide upon was that of a Linear Regression. While not the most complex model available at our disposal and therefore limited in scope, we selected this as it has the benefit of successfully captured and encapsulating our goal best of creating a correlation between salary and batting success.

Segment 2 Checklist (Presentation)

Outline the project, including the following

- Selected topic - Done
- Reason topic was selected - Done
- Description of the source of data - Done
- Questions the team hopes to answer with the data - Done
- Description of the data exploration phase of the project - Done
- Description of the analysis phase of the project - Done
- Presentations are drafted in Google Slides - Done

Segment 2 Checklist (GitHub Repository)

Main Branch

- All code necessary to perform exploratory analysis - Done
- Some code necessary to complete the machine learning portion of project - Done

README.md

- Description of the communication protocols - Done
- Outline of the Project - Done

Individual Branches

- At least one branch for each team member - Done
- Each team member has at least four commits for the duration of the second segment - Done

Segment 2 Checklist (Machine Learning Model)

- Description of preliminary data preprocessing
- Description of preliminary feature engineering and preliminary feature selection, including the decision-making process
- Description of how data was split into training and testing sets
- Explanation of model choice, including limitations and benefits

Segment 2 Checklist (Database Integration)

- Database stores static data for use during the project - Done
- Database interfaces with the project in some format - Done
- Includes at least two tables - Done
- Includes at least one join using the database language - Done
- Includes at least one connection string - Done

Segment 2 Checklist (Dashboard)

- Storyboard on a Google Slide(s) - Done
- Description of the tool(s) that will be used to create the final dashboard - Done
- Description of interactive element(s) - Done

Baseball Dataset Link

https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball

SQL lite schema: https://github.com/benhamner/baseball/blob/master/src/import.sql

CSV Tables Needed

-batting.csv https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball?select=pitching.csv

-salary.csv

https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball?select=salary.csv

-team.csv

https://www.kaggle.com/datasets/seanlahman/the-history-of-baseball?select=team.csv

Things to do

- Decide on a topic, source data, and perform exploratory data analysis. - Done
- Create a repository and establish individual branches for each team member. - Done
- Create a mockup of a machine learning model. - Done
- Create a mockup of a database - Done
- Decide which technologies will be used. - Done

Assignments (Segment 2)

- **Team:** Continue with analysis: add, commit, push, create new branches as needed, and utilize GitHub's built-in tools, such as PRs, to review the work you and your teammates have completed (this is an ongoing process, so keep it up!).
- **Square (Peyton):** Refine the machine learning model you'll be using (train and test).
- **Triangle (Jose):** Transform the mockup database into a full database that integrates with your work.
- **Circle (David):** Continue with analysis and create visuals to accompany the data story.
- **X (Charles):** Outline and begin work on a dashboard to house your final project. Check and test the work completed against the rubric.

Segment 3 Checklist - Presentation

Presentation tells story about project and includes

- Selected Topic - Done
- Reason topic was selected - Done
- Description of the source of data - Done
- Questions the team hopes to answer with the data - Done
- Description of the data exploration phase of the project - Done
- Description of the analysis phase of the project - Done
- Technologies, languages, tools and algorithms used throughout the project - Done

Segment 3 Checklist - Github Repository

Main

- All code necessary to perform exploratory analysis - Done
- Most code necessary to complete the machine learning portion of the project - Done

README.md

- Description of the communication protocols has been removed - Done
- Cohesive, structured outline of the project (this may include images, but they should be easy to follow and digest) - Done
- Link to Google Slides draft presentation - Done

Individual Branches

-At least one branch for each team member - Done

- Each team member has at least four commits for the duration of the third segment (12 total commits per person) - Done

Segment 3 Checklist - Machine Learning Model

✓ Description of data preprocessing - Done

✓ Description of feature engineering and the feature selection, including their decision making process - Done

✓ Description of how data was split into training and testing sets - Done

✓ Explanation of model choice, including limitations and benefits - Done for new and old model

✓ Explanation of changes in model choice (if changes occurred between the Segment 2 and Segment 3 deliverables) - Done

✓ Description of how they have trained the model thus far, and any additional training that will take place - Done

✓ Description of current accuracy score Additionally, the model obviously addresses the question or problem the team is solving. - Done

Information can be found here https://github.com/rclascano14/PROJECT/blob/main/ML_Information.md

Segment 3 Checklist - Dashboard

The dashboard presents a data story that is logical and easy to follow for someone unfamiliar with the topic. It includes the following:

- Images from the initial analysis - Done
- Data (images or report) from the machine learning task - ?
- At least one interactive element - Done

https://public.tableau.com/shared/66SKCF3YW?:display_count=n&:origin=viz_share_link