Predicting Hoefnagel Member Puzzle Hold Times

Jonathan Alexander, Madalyn Li, Hannah Luebbering, Aishwarya Singh

INTRODUCTION

Our goal is to build a model that can predict member's puzzle hold times given a puzzle's difficulty, piece count, and brand name. The predictions will help improve Hoefnagel Puzzle Club's current accuracy in determining the next optimal shipping location for a puzzle which will ultimately help reduce shipping costs and optimize customer satisfaction.

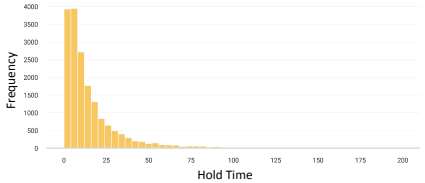
DATA OVERVIEW:

- **1059** Unique Puzzle Packs & **717** Unique Members
- **81.2%** of Puzzle Packs with 2 puzzles
- **51%** of Puzzle Packs with Average Difficulty Level
- **18.4** was the average member puzzle hold time
- **23** Member hold times greater than 200 days and less than 0.1 days (outliers)
- **5%** of the Hold Times with missing Puzzle Pack information

DATA ISSUES

- **139** Puzzle Packs with Missing Information
  Brand names were extracted from puzzle pack titles. The values of other attributes, such as piece count and difficulty, were estimated using historical data from each brand.
- **3.2%** Outliers in the member's data
  These data points were removed from the table

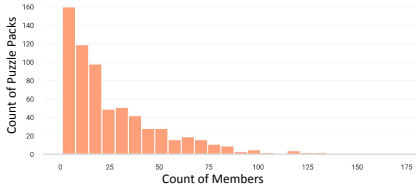
Figure 1: Distribution of Hold Time



MODELING CHALLENGES

- **92.8%** of Members had <= 75 historical data points
  This impacted our overall model performance because it is difficult to predict future hold times with limited historical information
- **Seasonality**
  Because we were unable to obtain relative days that members held a puzzle, we believe this impacted the model's ability to accurately predict hold times based on the time of year.

Figure 2: Puzzle Pack per Member



METHODOLOGY

- Conduct preliminary data analysis to better understand and pre-process the data. Identified possible patterns in pack names, analyzed the distribution of hold time, and looked for any apparent relationship between difficulty level and piece count of a puzzle pack
- Pre-process the data by imputing missing values and dealing with outliers based on the findings from data analysis
- Perform feature engineering, scaling, and column encoding to prepare the data for model building stage
- Build a model able to accurately predict member puzzle hold times given a puzzle's difficulty, piece count, and brand
- Evaluate the model performance on a random test set and compare the error rate against the benchmark model. Visualize the results and share insights

APPROACH

We observed that the predictive models performed better for members with more than 75 data points while the results were not as promising for new members. Thus, we designed a pipeline that uses a default model for new users and a personalized model for known users.

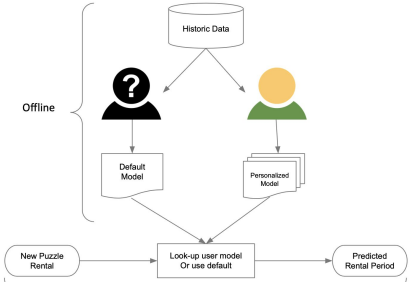


Figure 3: Live Pipeline

FEATURE IMPORTANCE

Some important features we identified in our model include:

Number of pieces for each difficulty level

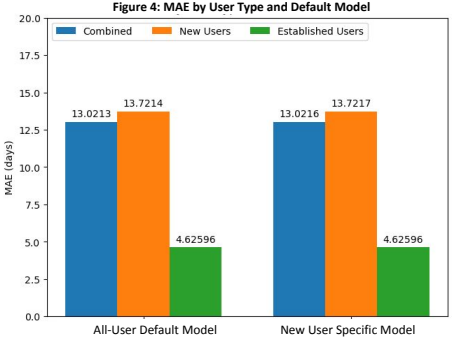
Avg hold times at member-level and pack-level

Brands and number of puzzles per pack

RESULTS

Our most accurate approach was attained using multiple linear regression models. For established users who had experience with puzzles customizing a model to them provided a significant decrease in the mean absolute error value. However, for users with little to no history on the platform our predictions remained quite inaccurate.

Figure 4: MAE by User Type and Default Model



NEXT STEPS

- Include seasonality as a factor in the model
- Deal with inconsistencies at the data extraction stage
- Try different approaches to handling data inconsistencies such as median member hold times
- Add more member hold time data to improve predictions
- Reduce the model's dependency on historical hold time data by adding regularization

ACKNOWLEDGEMENTS: We'd like to thank the Hoefnagel Wooden Jigsaw Puzzle Club, our sponsor Maya Gupta, our professor Megan Hazen, and the UW MSDS program