DATA601 Presentation

Instructions

9-10 minutes speaking
5 minutes answering Qs

Intro

Name Hi my name is Samuel Love
Title My project title is From Bricks to Bucks: Exploring Influential Variables in Sales Patterns of Lego Bricks
Provided My project was provided by UC Maths & Stats department
Lego description Lego is a worldwide brand of plastic

construction toys produced by Danish company The Lego Group since 1949

Project descriptions The project brief had a wide range of potential questions, so I had freedom to choose my goals and how to achieve them

Aims

Main Research Q 'Can we accurately predict the secondary prices of Lego sets using a defined set of characteristics?'

Secondary Research Q 'Do Lego sets based on licensed intellectual properties differ from Lego original sets?'

Specifics I examined trends in Lego set characteristics over time and trained several regression machine learning algorithms to generate models for predicting secondary market prices

Software I used was the Shiny package from R Studio to create an interactive dashboard of visualisations

Objectives

5 main objectives

Time-series analysis based on the year that sets were

released
Visualisation to help with data cleaning and for analysis of time-series data, clustering effects, and model metrics
Clustering k-means clustering on 2 standardised subsets of data to explore any organic groups in the data
Predictive modelling using 4 algorithms linear regression (LR), elastic net (EN), random forest (RF), and support vector machine (SVM) as well as a null model
Feature ranking performed by EN and RF primarily to see if historical price data is useful good predictor of current prices

Data Source

I used **4 datasets from 3 sources**, all csv files
Main dataset JB contains 19,409 observations of 36 variables sourced from the website Brickset.com via its API
... notable variables include theme, year released, number of pieces
... there is recommended retail price but no secondary market prices
Secondary datasets D15 and D18 contain 2,322 observations of 11 and 22 variables respectively
... from Mendeley Data which is a research database **...** same notable variables but also includes average yearly secondary

prices for 2015, 2018, 2019
Secondary dataset MP contains 5,075 observations of 9 variables
... someone personal project from Kaggle
... same notable variables but also includes average yearly secondary prices for 2022
Key variable used for joining is setID

Data Limitations

There are many **limitations** to consider when using this data
Biased data for example sampling bias (from popular sets being overrepresented) or selection bias (from data collectors choosing which sets to track)
Lack of context especially with MP as there is limited accompanying documentation on collection process
... JB has thorough API documentation and D15 and D18 have an accompaning research paper
Incomplete data 2015 prices are given as yearly averages, but 2018 and 2019 are monthly with 2019 being only the first four months of the year
Small sample size as D15 and D18 are 1/8th the size of JB whilst MP is 1/4
Assumptions include $USD, quality of sets being as-new, consistent characteristic variables

Changing landscape which this static data does not represent
... There could be effects from seasons, popularity, or which sets are retiring soon

Method - Visualisations

Steps for producing visualisations are broadly cleaning, joining, clustering, and visualising the data
Visualisations are used at every step
Specific plots include Histograms, boxplots, and homogeneity plots for data cleaning
... bar graphs and scatterplots of characterisitc variables, with extras for comparing price with other variables
... line graphs of the elbow and silhouette methods to determine optimal cluster number
... pairwise plot and bar graphs to determine effects of clustering

Method- Machine Learning

Steps for producing effective machine learning models are broadly subsetting, training, comparing metrics, and testing

the best performing models

Subsets reassemble the supplementary datasets but have no missingness and are standardised

... the goal is to use supervised learning to predict prices for 2019 and 2022 respectively

... the 2019 subset has 2,277 observations of 12 variables **...** the 2022 subset has 3,241 observations of 7 variables

... categorical variables theme and licensed were one-hot-encoded, then all variables were centred and scaled

... comparisons focus on the effects of price variables RRP, 2015, and 2018 which are present only in the 2019 subset

Training I used a 75/25 train test split and created 20 bootstraps of the training data

Metrics I chose three metrics suitable for evaluating regression model performance

... R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables

... Root Mean Square Error measures the standard deviation of the residuals

... Mean Absolute Error measures the average absolute difference between the observed outcomes and the predictions

Test All models performed better than null, with elastic net and linear regression being nearly identical

... I therefore tested the elastic net, random forest, and support vector machine models

Results - Visualisation

There are 5 main **results** from visualising the data
There is an **increasing number of sets** released each year
The characteristics of sets **increased in diversity** especially in later years
The average secondary market **prices consistently varied** depending on the year sets were released
More **complex sets** had higher prices
There is some feasible **organic clustering** of price variables
Boxplots of prices categorised by **licensing**, proved inconclusive

Results - Predicting 2019 Prices

When predicting 2019 prices SVM had the best average R^2 0.31 and RMSE of 0.92
... whilst EN had the best MAE of 0.08
The most **significant predictor** was 2018 prices with others being RRP, 2015 prices, cluster and number of pieces

Results - Predicting 2022 Prices

When predicting 2022 prices EN had the best average R^2 of 0.62 and RMSE 0.62
... whilst SVM had the best MAE of 0.21
The most **significant predictor** was cluster with others being pieces and year released

RF performed better in training than testing

Conclusions

Based on this data, Lego sets are **increasing in quantity and complexity** over time
There is **high variation in average prices**
Recent historical **price variables are great predictors** for current prices.
There was no optimal model however, there were marked differences in performance metrics
... 2022 models had much improved R^2 and RMSE metrics with slightly worse MAE

To answer my research questions, there are minor differences between licensed sets and original sets, however it is difficult

to parse, especially considering the high proportion of themes that contain both licensed and original sets.
It did prove possible to produce models that perform better than null for accurately predicting secondary prices of Lego sets using a defined set of characteristics
Future work could involve more robust datasets and the use of ensamble methods to improve the predictive power of the models

Wrap up

Thank you for listening. Thanks Paul and Phil for supervising my project. Any questions?