Video Game Sales Prediction

Final Report

Boya Zhang, Jiaxuan Wu

Overview

- **Goal:** Predict first year global sales
- **Source**
  - Kaggle(https://www.kaggle.com/ashaheedq/video-games-sales-2019)
- **Time Range:** 2010 - 2019
- **Data Size** (after pre-processing)
  - Around 7000 records, 4000 games
- **Features**
  - Categorical: Genre, ESRB Rating, Publisher, Platform
  - Numerical:
    - Critic Scores: missing value - new feature: if_not_scored
    - Price: missing value imputation - genre platform average

Data Analysis Exploration

- **Critic Score**
  - Add a new column to show whether the critic score is missing or not
  - Reason: A missing of critic score might means this game do not have enough popularity.
- **Price**
  - Log transformation
- **Global Sales**
  - Reset outliers that is larger than 95% percentile value to the 95% percentile value
  - Log transformation

Model

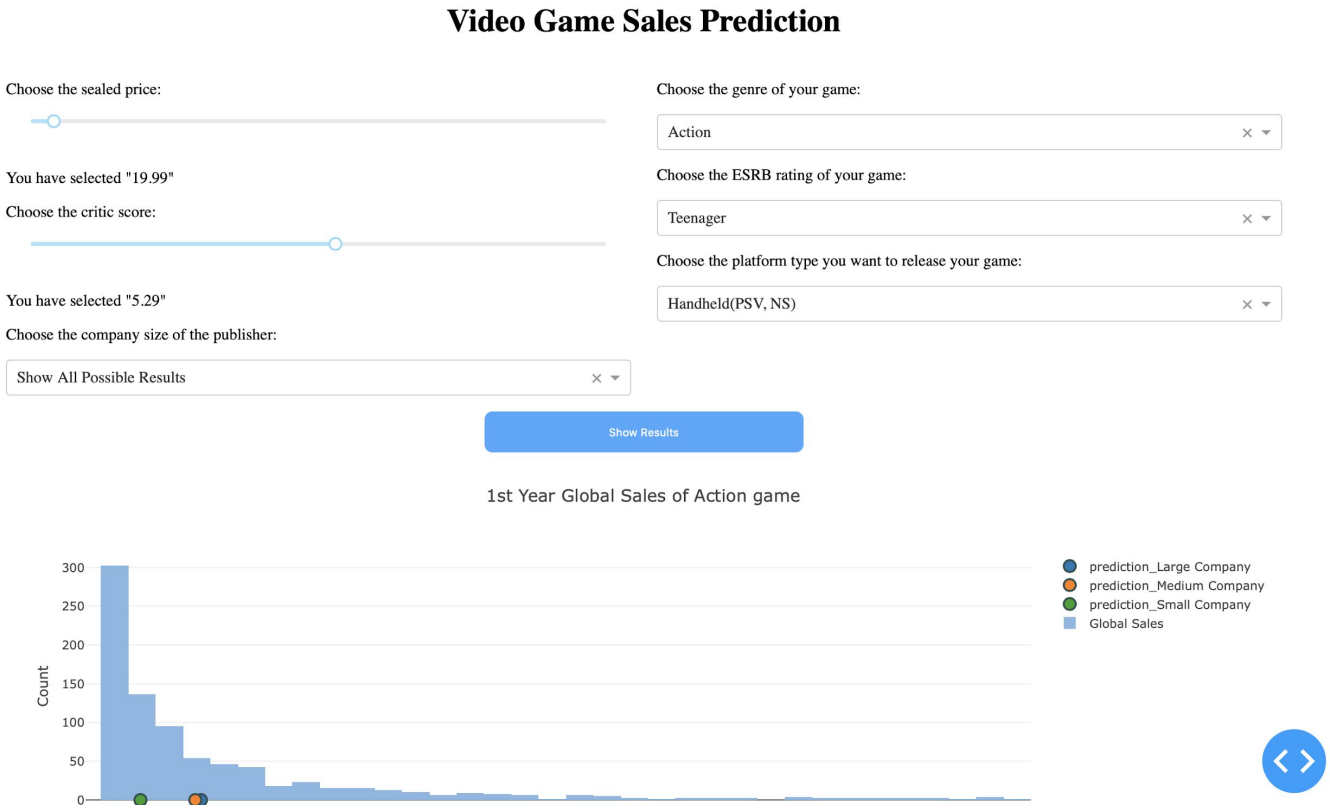
- **80% training set, 20% test set**
- **Baseline Model --- Linear Regression**
  - Evaluation result: R square --- 0.46, MSE --- 1.17
- **Random Forest Regression**
  - 5-fold cross validation to tune parameters: # trees, max_depth
  - Optimized result (# trees = 200 , depth = 8)
  - Evaluation result: R square --- 0.47, MSE --- 1.16

Attempts Result

		Without Grouping	Attempt 1	Attempt 2 (Last Week)	Attempt 3
Baseline	R square	0.24	0.38	0.43	0.46
	MSE	1.62	1.6	1.5	1.17
Random Forest	R square	0.31	0.44	0.47	0.47
	MSE	1.46	1.46	1.41	1.16

- **Without Grouping**: do not group platform, ESRB rating
- **Attempt1**: Group platform by company, console type, ESRB rating by 12-, 17-, 17+, Pending/Missing
- **Attempt2**: Regroup platform only by console type
- **Attempt3**: Add data pre-processing step mentioned in previous slides.

Visualization



Q & A