Studio CV:

Predicting Anime Sales via Linear Regression

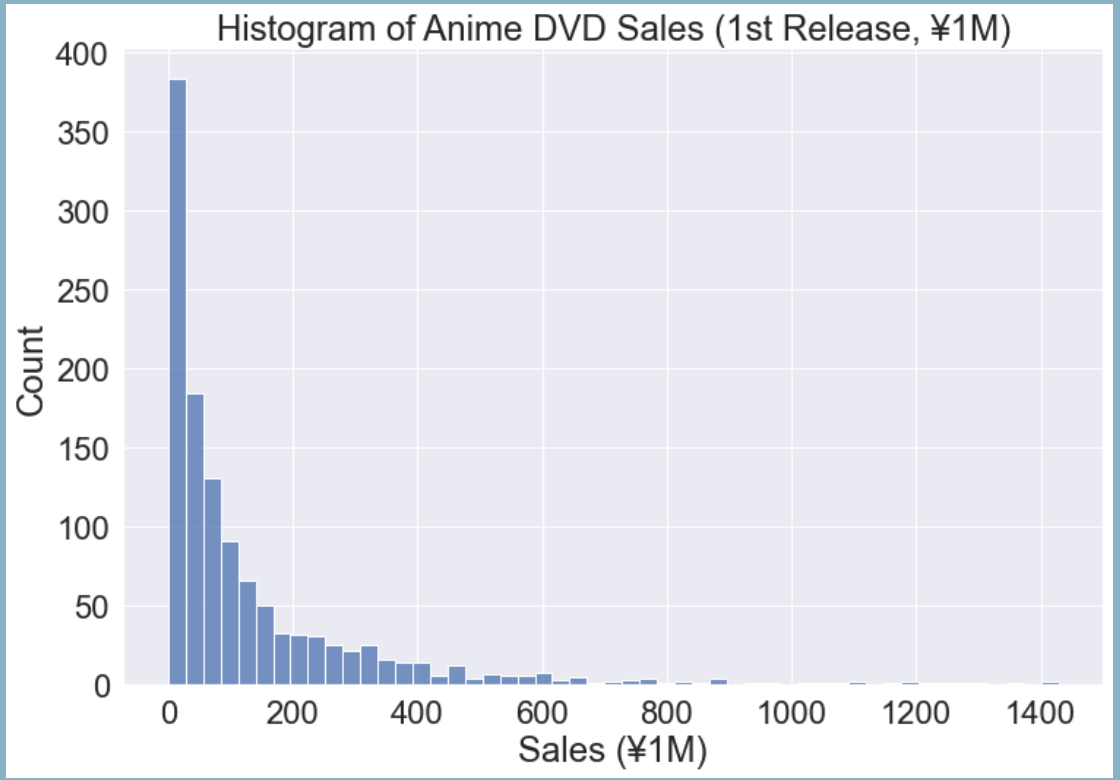
Modeling by:
Andrew Zhou

Background, Objectives, & Data Sources

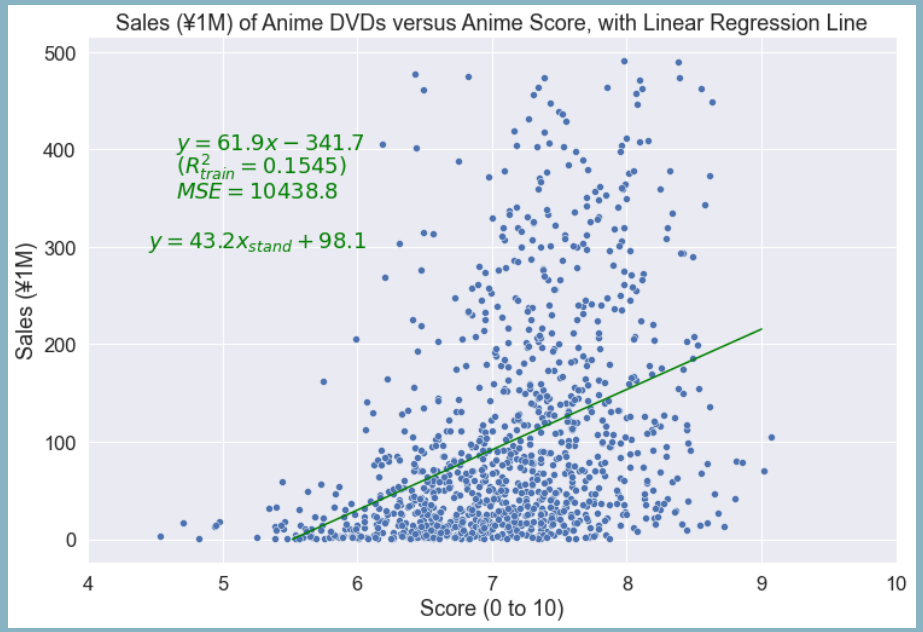
- DVD Sales: an important proxy for overall “success”
- Most visible metric; devoted fans tend to obsessively monitor
- Build a model and identify key features associated with anime sales

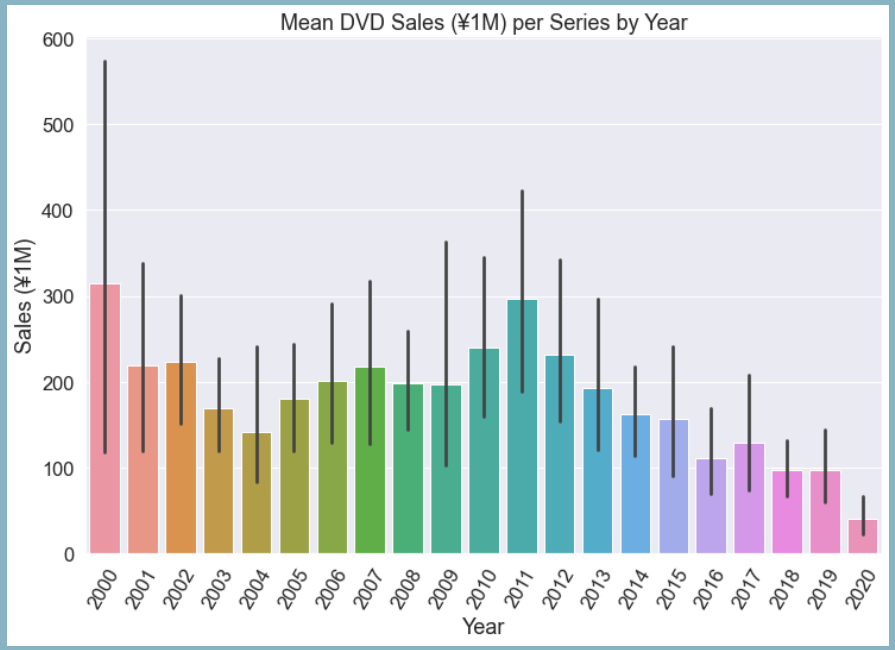
Sales Data
https://www.someanithing.com/

Series Data
https://myanimelist.net/

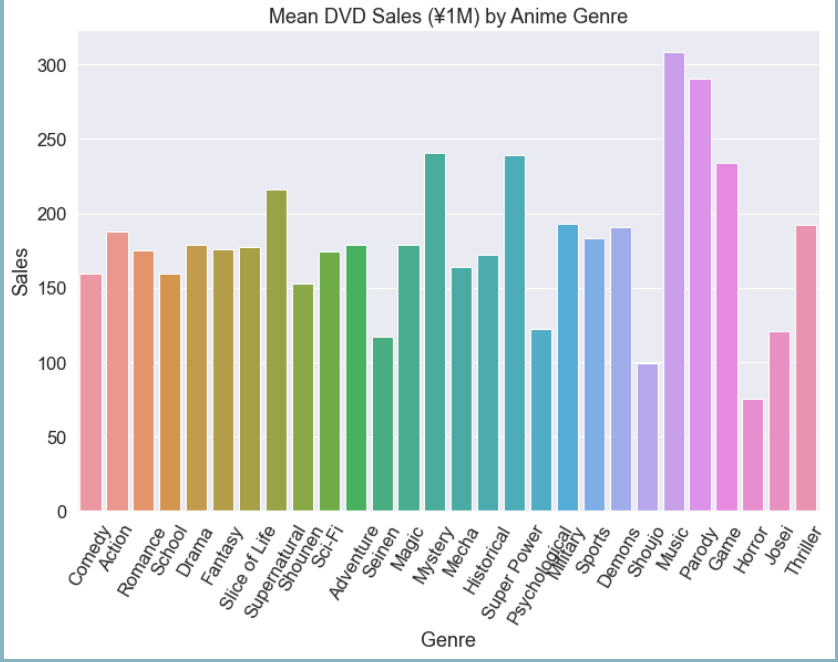


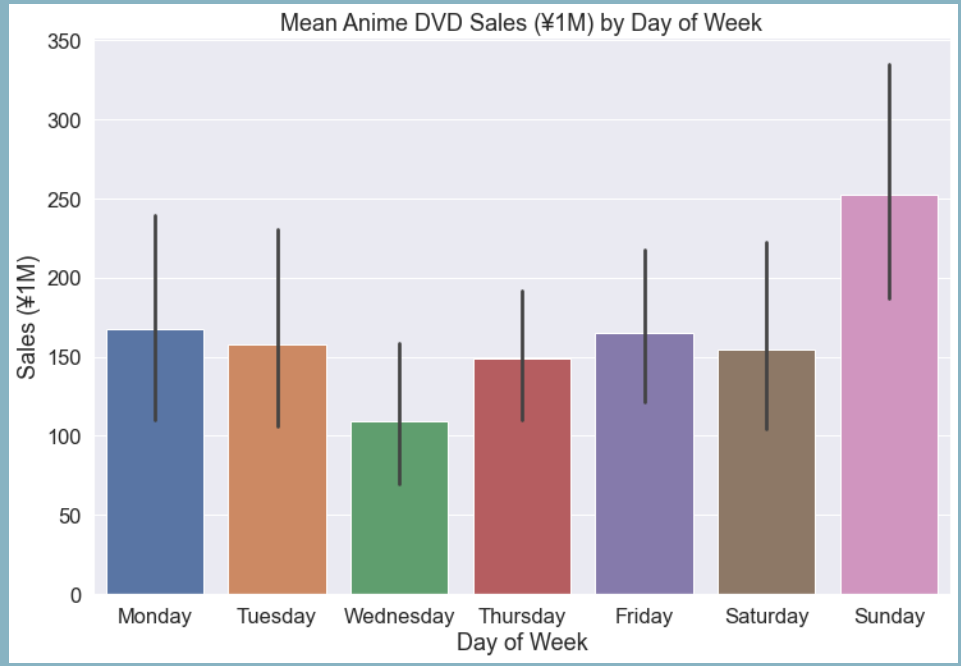
Base Features: Score, Year





Base Features: Genre, Day of Week

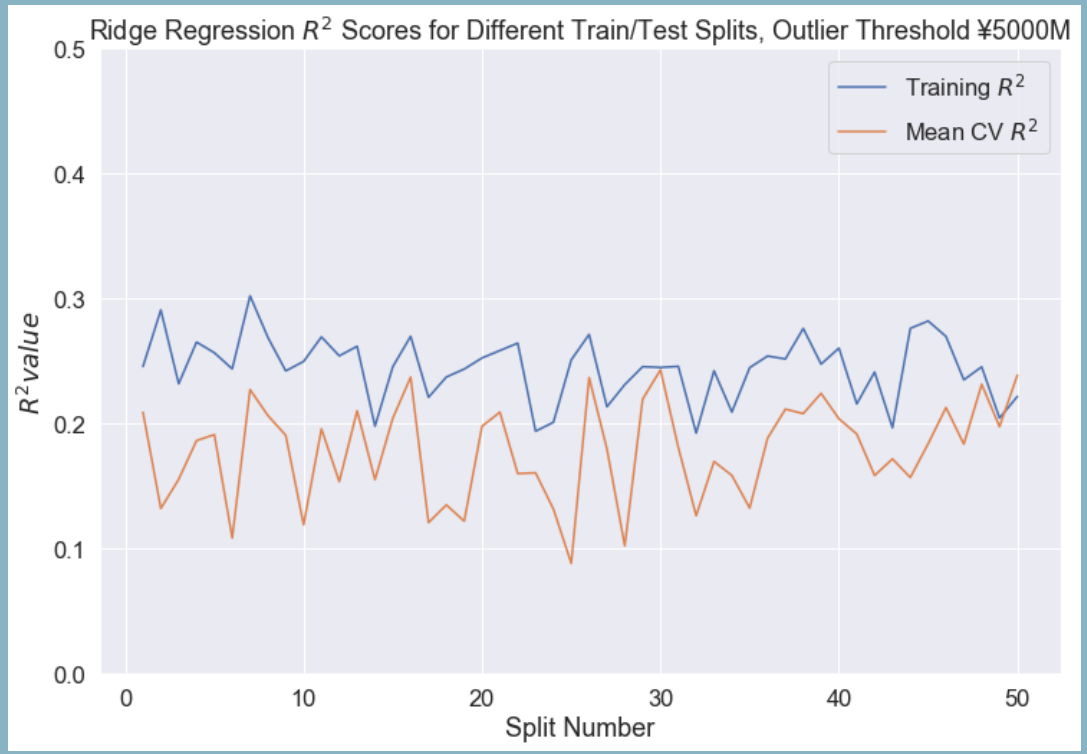




Removing Outliers: Numerical Instability

- There exists a long tail of high sales values
- Leads to numerical instability in our models
- Generate toy Ridge models for different train/test splits

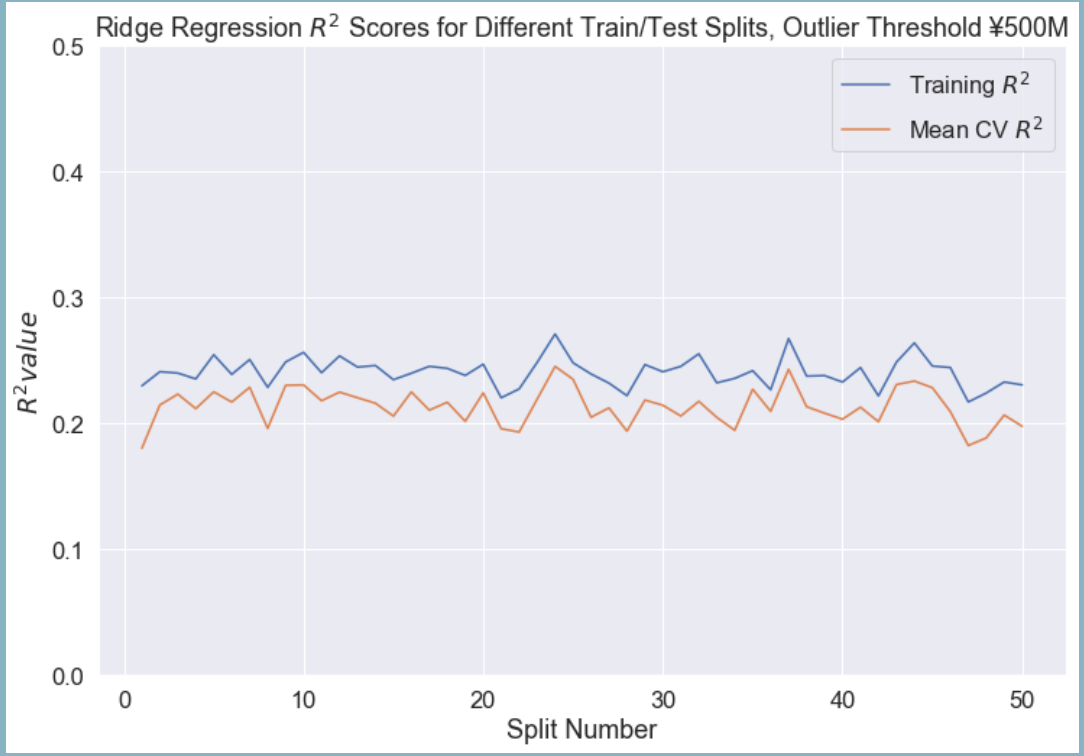
Quantity	μ	σ	Approx. RMSE
$Sales$	¥164M	¥311M	—
R^2_{train}	0.245	0.0253	¥270.3M
$\mu(R^2_{CV})$	0.179	0.0393	¥281.9M



Removing Outliers: Stability Achieved

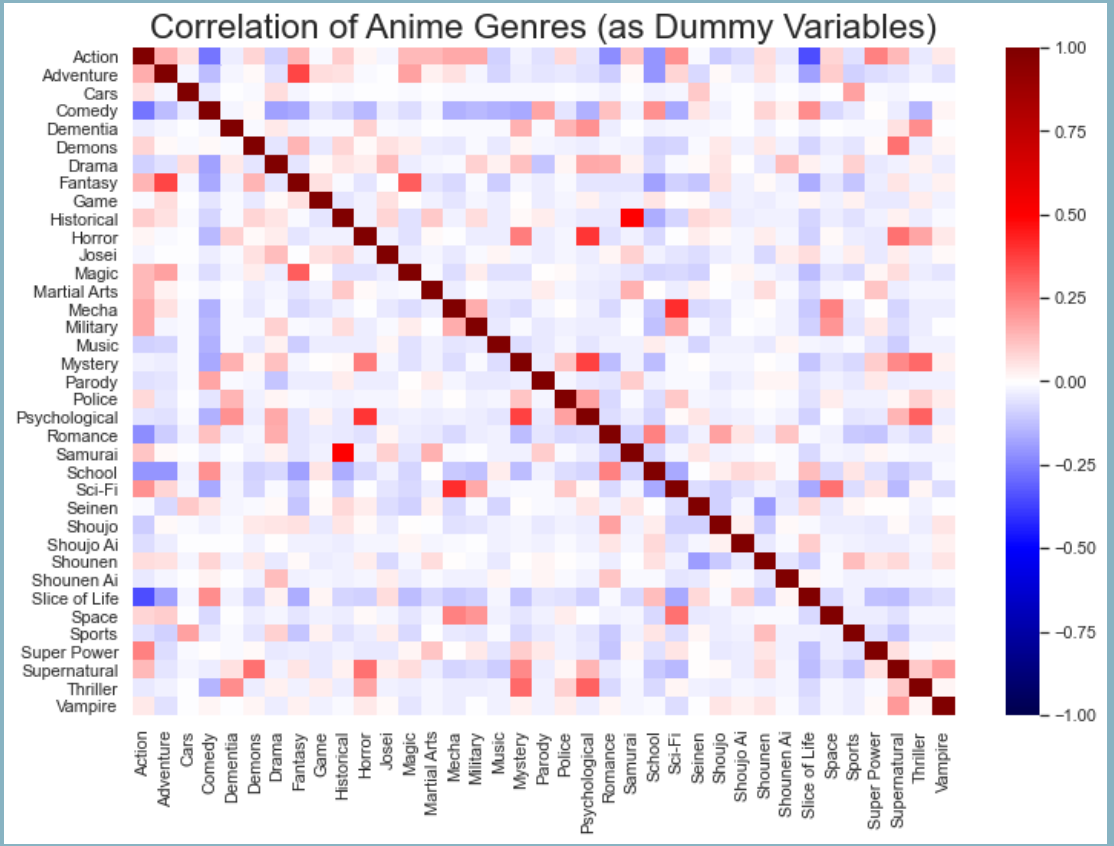
- Remove 85 outliers with sales above ¥500M
- Vastly more stable
- Both variance and value of R2 are significantly lower

Quantity	μ	σ	Approx. RMSE
$Sales$	¥99.8M	¥111.1M	—
R^2_{train}	0.241	0.0115	¥96.8M
$\mu(R^2_{CV})$	0.213	0.0145	¥98.6M



Feature Engineering: The Problem of Genre

- Genres are correlated, and each show can have many
- Massive number of features, especially including interaction & polynomial terms
- Overfitting & uninterpretability



Feature Engineering: Genre Clustering by K-Means

Cluster No.	Name	Genre 1 (Value)	Genre 2 (Value)	Genre 3 (Value)
1	Fantasy Adventure	Fantasy (0.988)	Adventure (0.595)	Action (0.571)
2	SoL Comedy	Comedy (0.840)	Slice of Life (0.587)	School (0.359)
3	Action	Action (0.994)	Comedy (0.489)	Supernatural (0.391)
4	Romcom	Romance (0.914)	Comedy (0.840)	School (0.691)
5	Romdram	Drama (0.966)	Romance (0.691)	Comedy (0.371)
6	Supernatural Mystery	Mystery (0.886)	Supernatural (0.647)	Drama (0.489)
7	Scifi	Sci-Fi (0.969)	Action (0.704)	Mecha (0.396)

Final Model: Features & Methodology

- Cross-Validated Lasso Regression with Polynomial Features (d=2)
- Strategy: input a large number of features and rely on lasso for selection

Feature Type	Count
Initial (cont. & cat.)	7
Pre-poly (cat. dummies)	34
Polynomial	630
Post-model, $\beta_i \neq 0$	44
Post-model, $\lvert\beta_i\rvert \geq 1$	23

Feature	Values
Score	0.0 to 10.0
Year	2000 to 2020
Rating	G, PG, PG-13, R, R+
Members	integer, 0 to approx. 2000000
Favorites	integer, 0 to approx. 150000
Timeslot	4-hour timeslot, e.g. 8PM-12AM
Day of the Week	Monday to Sunday
Genre Cluster	One of 7
Polynomials	Degree ≤ 2, incl. interactions

Final Model: Evaluation

Model Results

Data Set	Sales Variance $(¥1M)^2$	R^2	MSE $(¥1M)^2$	RMSE $(¥1M)$
Train	9166.79	0.292	6457.34	80.36
Test	9060.28	0.255	6749.77	82.15
Mean CV	—	—	7188.02	84.78

Top-Weighted Features

Feature 1	Feature 2	Coefficient
Score	—	25.60
Year	—	-18.54
Members	—	10.91
4-8PM	Romcom	7.90
Favorites	Thursday	4.33
Favorites	Wednesday	3.59
G Rating	Scifi	2.96
Friday	Romdram	-2.76
Members	R+ Rating	2.54

Future Work

Feature Engineering

- Address overfitting
- Optimize genre clusters
- Reduce number of categories
- p-value analysis

Broader Questions

- Extensions and real-world applications?
- Interpretability and collinearity
- Sales: proxy for “success”?

Thank You!

Sources

Data

https://www.someanithing.com/
https://myanimelist.net/

Images

https://wallpaperaccess.com/minimalist-totoro
https://www.vulture.com/2019/10/studio-ghibli-movies-streaming-hbo-max.html