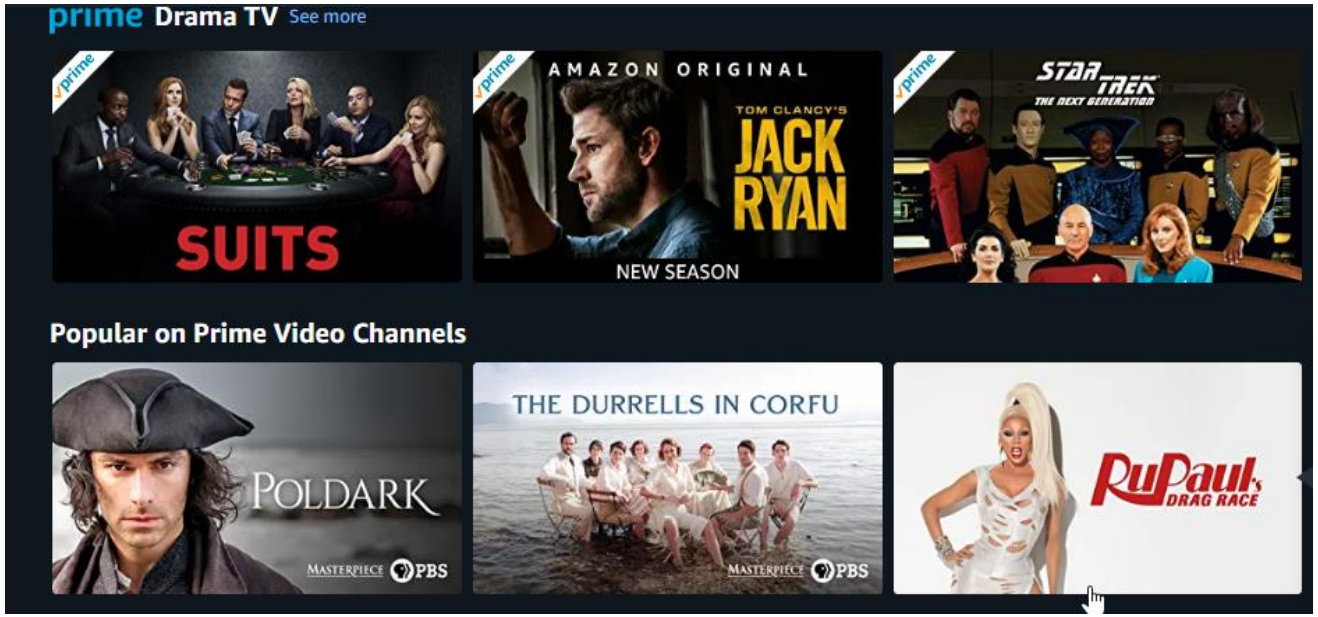
Amazon Prime Video View Time per Day Prediction

Data Incubator Project

11/20/2019

Fei Yu

Better Recommendations for Videos



- Two benefits of making better video recommendations to customers:
  - Better business value
  - Better customer relationship

- One reliable quantitative metric: Cumulated View Time (CVT) per Day

- Objective: use machine learning to predict CVT for videos

Content

- Amazon Video Data
- Data Preparation
- Model Evaluation
- Feature Importance Analysis
- Summary

Amazon Video Data

Dataset: 4,226 rows, 16 columns. Each row is one movie.

	A	B	C	D	E	F	G
1	video_id	cvt_per_day	weighted_categorical_position	weighted_horizontal_poition	import_id	release_year	genres
2	385504	307127.6056	1	3	lionsgate	2013	Action,Thr
3	300175	270338.4264	1	3	lionsgate	2013	Comedy,C
4	361899	256165.8674	1	3	other	2012	Crime,Dra
5	308314	196622.721	3	4	lionsgate	2008	Thriller,Dr
6	307201	159841.6521	1	3	lionsgate	2013	Crime,Thri
7	389496	135076.6098	1	5	mgm	2000	Comedy
8	385507	134155.7402	1	6	lionsgate	2013	Action,Adv
9	380517	116906.0079	1	7	lionsgate	2014	Western,D
10	369857	116871.1216	2	9	lionsgate	2013	Thriller,Cri
11	393463	111565.5967	2	7	lionsgate	2009	Action,Adv

imdb_votes	budget	boxoffice	imdb_rati	duration_	metacritic	awards	mpaa	star_category
69614	15000000	42930462	6.5	112.301	51	other	PG-13	1.71
46705	15000000	3301046	6.5	94.98325	41	no	R	3.25
197596	26000000	37397291	7.3	115.7637	58	other	R	2.646666667
356339	15000000	15700000	7.6	130.7036	94	Oscar	1.666667	
46720	27220000	8551228	6.4	105.5455	37	other	R	3.066666666
13250	60000000	32095318	5.5	98.46835	37	no	PG-13	2.75
16188	11000000	8551228	5.2	94.33642	57	other	R	2.74

➢ **Prediction:**

CVT – Cumulated View Time [Seconds] per day *(indicates whether a movie is popular or not, more profit)*

➢ **Feature examples**

- Weighted_vertical_position
- Weighted_horizontal_position
- Genres (Drama, comedy,…)
- Release_year (Year 1920 ~ 2019)
- IMDB_votes (total IMDB votes)
- IMDB rating (0~10)
- Duration_in_minutes (mins)
- Budget ($)
- Awards (Oscar, other award)
- MPAA (PG-13, R, …)
- Boxoffice

➢ **Supervised learning**

➢ **Three Key Questions:**

1. Predict CVT of a new movie based on its existing features
2. Which features are important to the CVT for a new movie

Help video website optimize website to do better recommendations

Data Preparation

- **Deal with missing data – used mean values**
  - For budget, missing data is not very obvious; no NAs; but zero budget indicates missing value

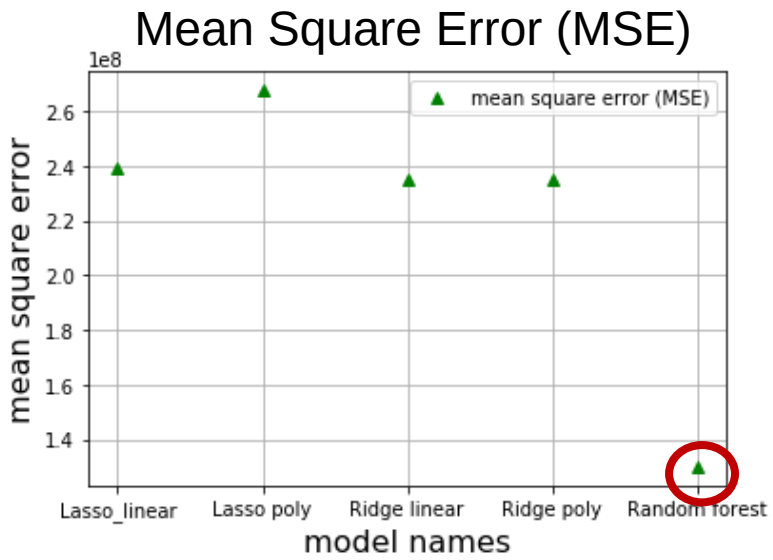
- **Feature Engineering**
  - Bin the release years for every 10 years (1920 ~ 2019)
  - Categorical features – one hot encoding
  - Feature standardization
    - Since different features are in different scales

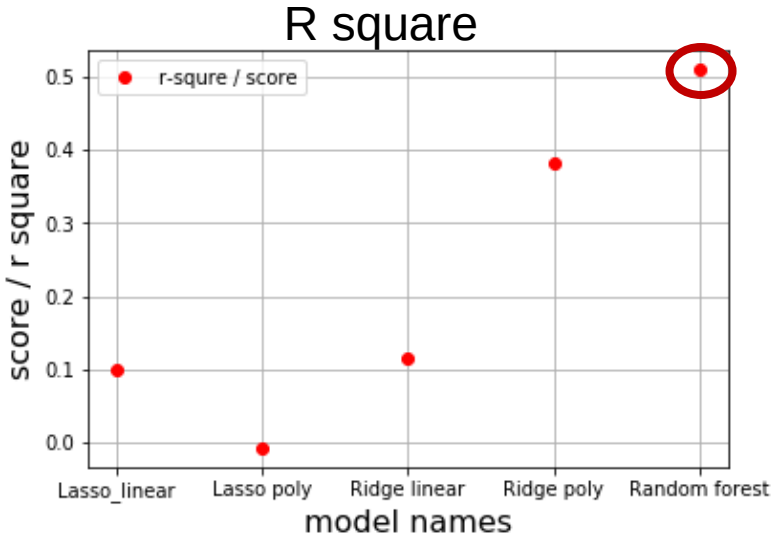
- **80% for Training & 20% for testing with cross-validation**

Model Evaluation

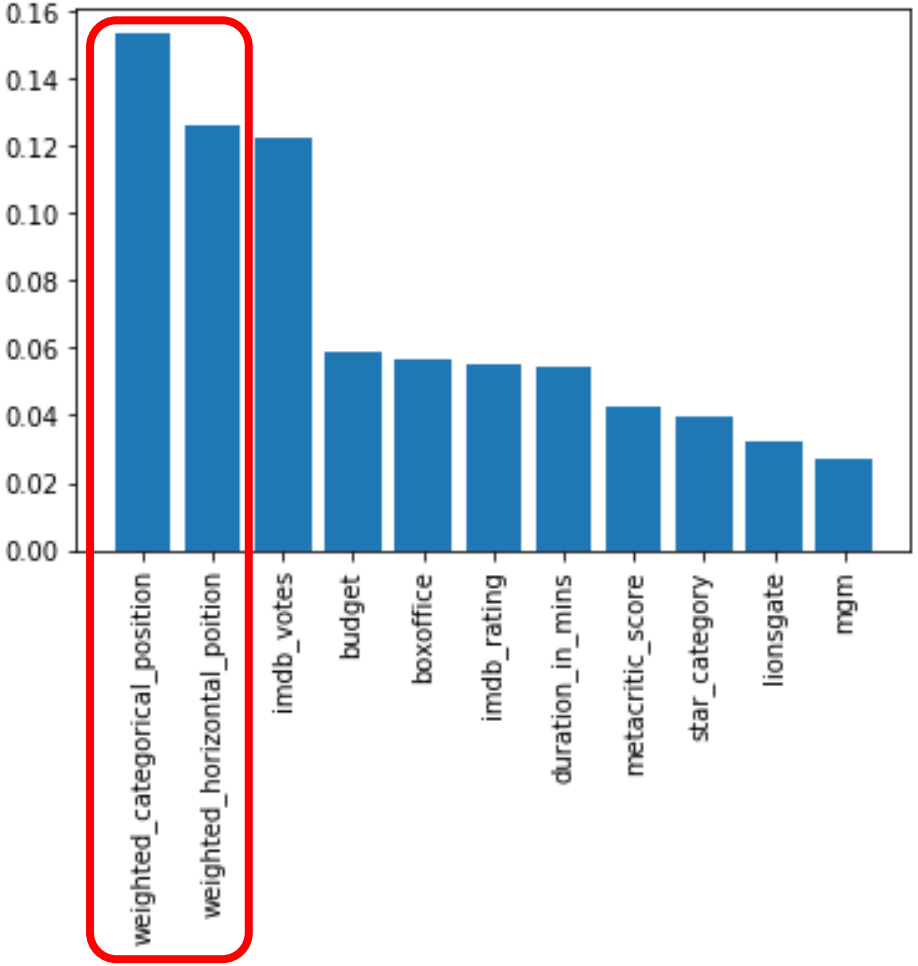
Random Forest model has the highest R Square, lowest MSE

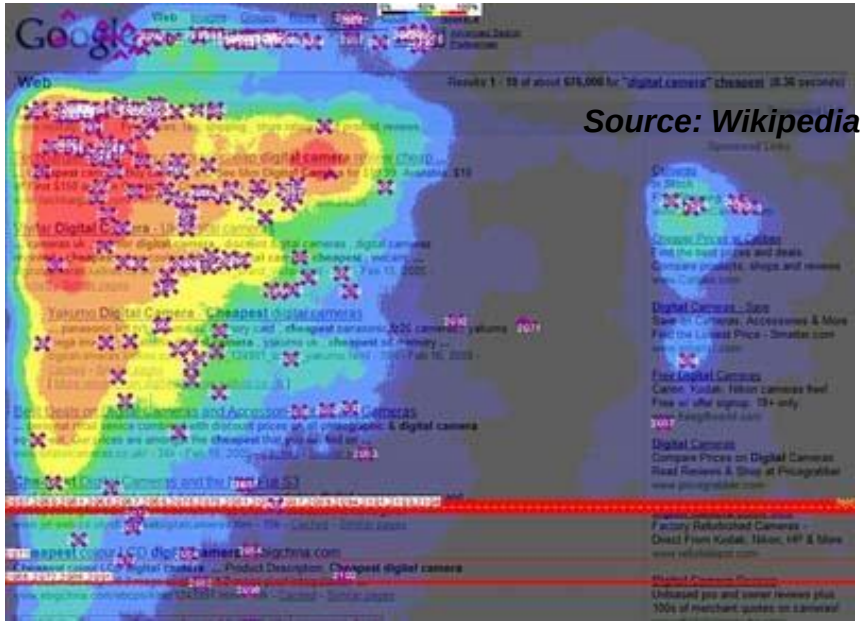
- **Modelling**
  - **Linear model**
    - All features with Lasso
    - All features with Ridge
    - Polynomial features with Lasso
    - Polynomial features with Ridge
  - **Nonlinear model**
    - Random forest





Feature Importance Analysis





Google Golden Triangle
https://www.mediapost.com/publications/article/235341/the-evolution-of-googles-golden-triangle.html

- Top two important features: **the categorical and horizontal positions** of video on the website
- Consistent with Google Golden Triangle Rule
- The least important feature: release year

Summary

✓ **Two Key Questions:**

1. Predict CVT of a new movie based on its existing features
2. Which features are important to the CVT for a new movie

✓ **Take-Aways:**

1. This model can help website manager to predict the CVT for new movies so to help them optimize movie recommendations

2. The location of movie is the most important feature for the CVT

3. This result also applies to other similar video websites or online shopping websites