Movie Advisor

Predicting upcoming movies’ box office revenue in Taiwan for theater manager to plan released weeks and halls for new movies .

Team 3

Sam Wang

Sean Xie

Jenny Wang

Jessica Deng

Business goal

Main stakeholder : Theater manager

Business problem : Since theater managers have to arrange released weeks and halls for a new movie, they have a potential need of knowing which new movie will make good box office revenues.

Data mining goal

To predict box office revenues in Taipei based on movie features and box office revenues in USA. Since we are interesting in which movie factors are more powerful to box office revenues, this will become an ongoing, predictive, and supervised task, and the main outcome variable of interest is box office revenue in Taipei.

Data description

Source: Yahoo! movie / Atmovies / PTT / True movie / IMDB / YouTube / Dorama

Row: A movie record

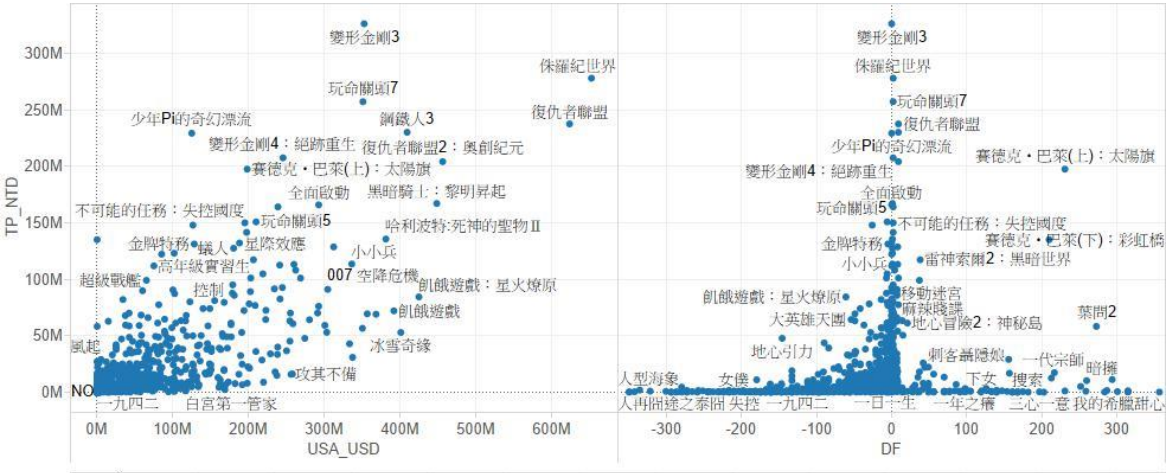
Name_CN	Name_EN	Date_TW	Length(raw)	Length(min)	Director	Cast	Agent	Expection	Type	Production	Country	Language	date_US	Budget_USD	USA_USD	IMDB	YouTube	DF
破碎的擁	Broken E	40186	0.09	128	《悄悄告	《情遇巴	山水	0.91	劇情/愛	Universal	Spain	Spanish	40137	1.8E+07	4930000	7.2	1828	-49
打不倒的	INVICTUS	40193	0.09	134	《登峰造	摩根費里	華納兄弟	0.94	劇情	Warner B	USA	English	40148	6E+07	3.7E+07	7.4	32940	-45
鼠來寶2	Alvin and	40199	0.06	89	《怪醫杜	《全民公	福斯	0.95	喜劇	Fox 2000	USA	English	40170	7.5E+07	2.2E+08	4.5	102997	-29
歐吉桑卡	Old Dogs	40200	0.06	87	《荒野大	《荒野大	博偉	0.95	喜劇	Walt Disr	USA	English	40142	3.5E+07	4.9E+07	5.4	2257	-58
帕納大師	The Imag	40200	0.08	122	《神鬼剋	《黑暗騎	威視電影	0.95	奇幻/劇	Infinity F	France	English	40172	3E+07	7670000	6.9	12153	-28
食破天驚	Cloudy W	40207	0.06	90		(配音)《	博偉	0.9	動畫	Columbia	USA	English	40074	1E+08	1.2E+08	7	15190	-133
墮落與智	FILTH & V	40207	0.06	84	瑪丹娜(M	尤金・赫	聯影／聯	0.74	劇情	Semtex F	UK	English	39747	5300000	22406	5.6	1525	-460
奪天書	The Book	40221	0.08	118	《開膛手	《亡命快	博偉	0.91	動作/劇	Alcon Ent	USA	English	40193	8E+07	9.5E+07	6.9	2599	-28
沙灘上的	The Beac	40235	0.08	110	安妮華達	(Agnès V	聯影／聯	0.62	歷史/傳	Ciné Tam	France	French	40034	4000000	239711	7.9	4452	-201
攻其不備	The Blind	40235	0.09	128	《心靈投	《愛情限	華納兄弟	0.96	劇情/溫	Alcon Ent	USA	English	40137	2.9E+07	2.6E+08	7.7	6922	-98

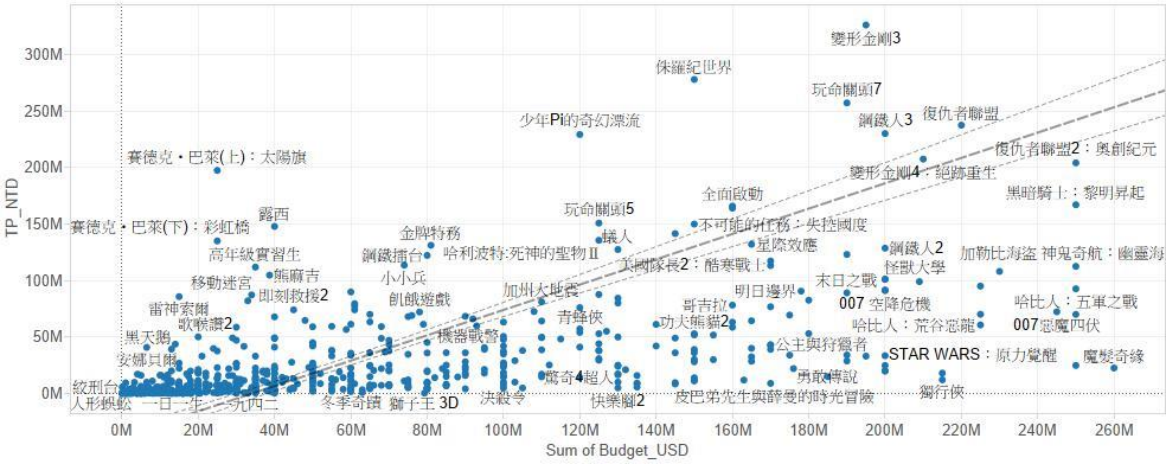
Caring	War	Drama	Action	Horror	Suspense	Love	Fiction	Adventure	Inspiration	Comedy	Fantacy	Animation	Crime	Documentar	Musical	History	Rating_G	Rating_PG	Rating_PG-1	Rating_R	TP_NTD
0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2370000
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4510000
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	8840000
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1.3E+07
0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1.3E+07
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1E+07
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	140000
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	9510000
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	410000
1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1.6E+07

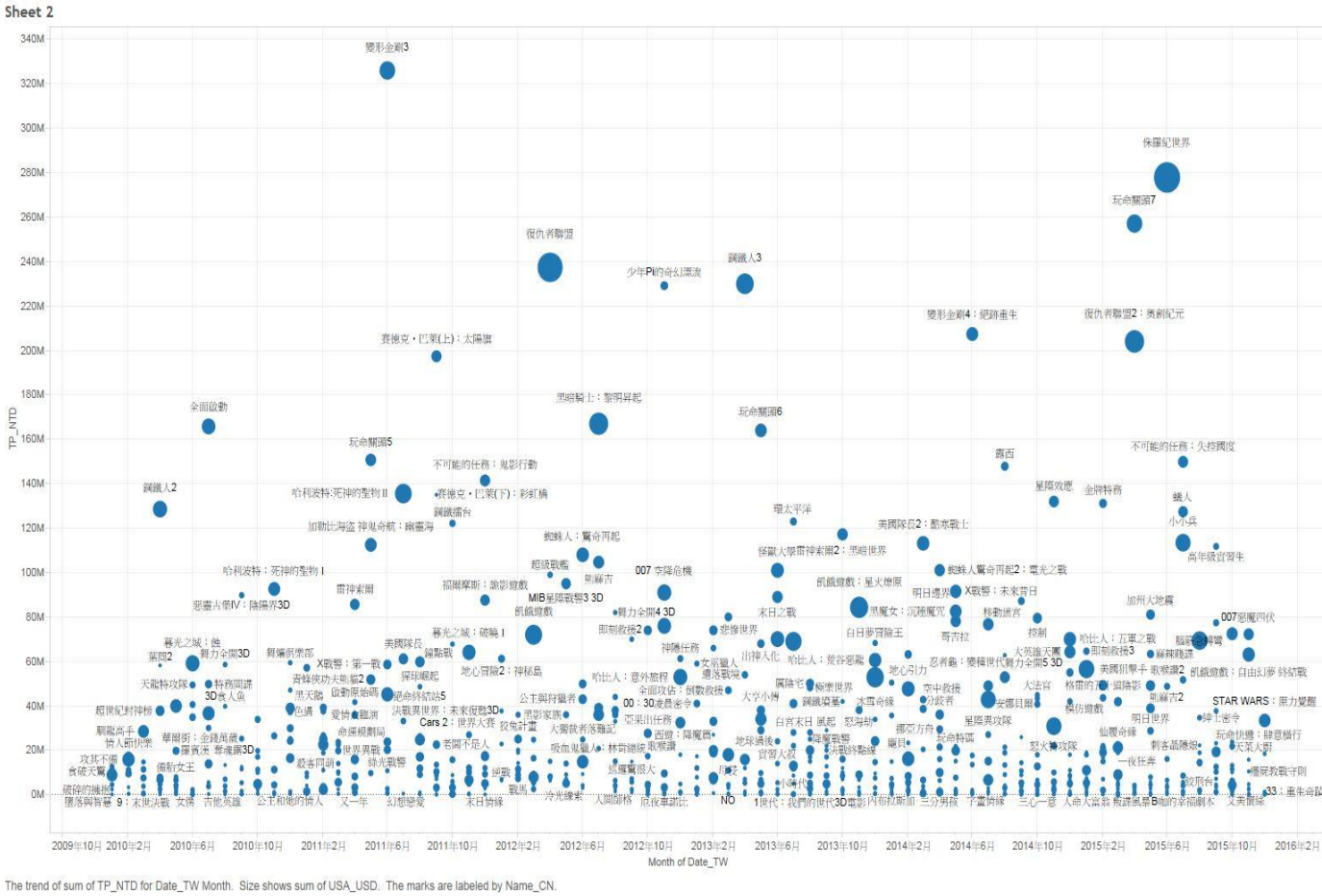
Data pre-processing

1. Create dummies for Rating / Type
2. Fill in missing Budget data by clustering and getting the median for each cluster
3. Remove movies that released earlier in Taiwan than in USA
4. Remove movies which just released

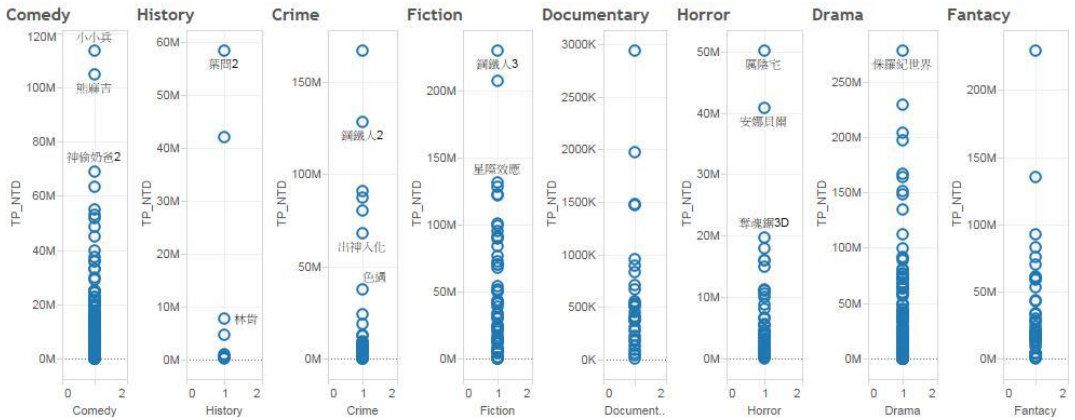
Data visualization

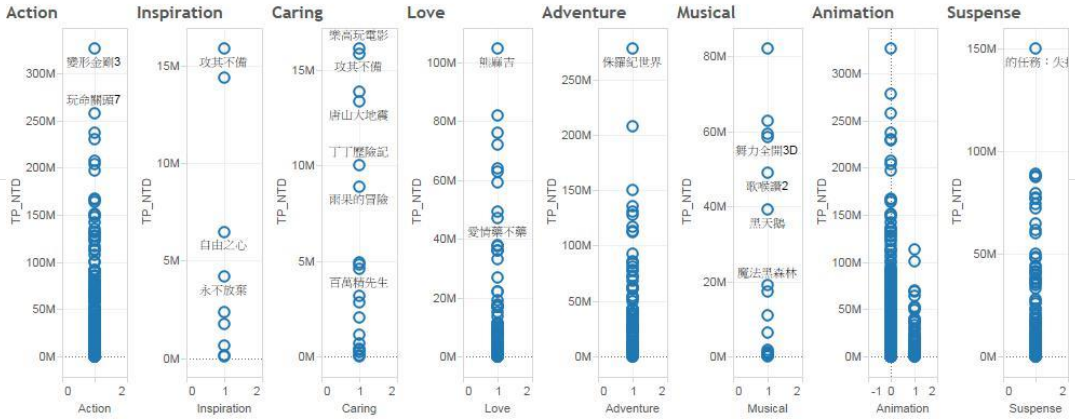




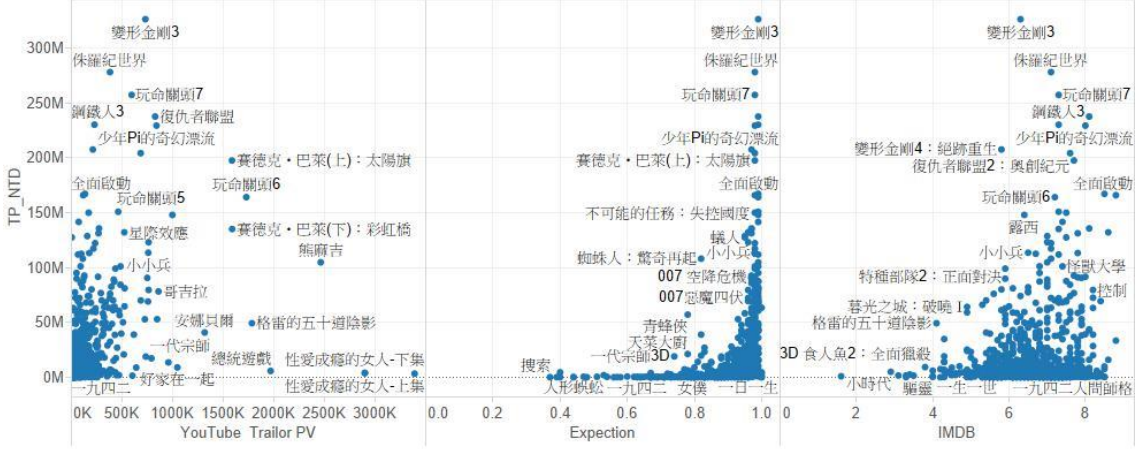


The trend of sum of TP_NTD for Date_TW Month. Size shows sum of USA_USD. The marks are labeled by Name_CN.

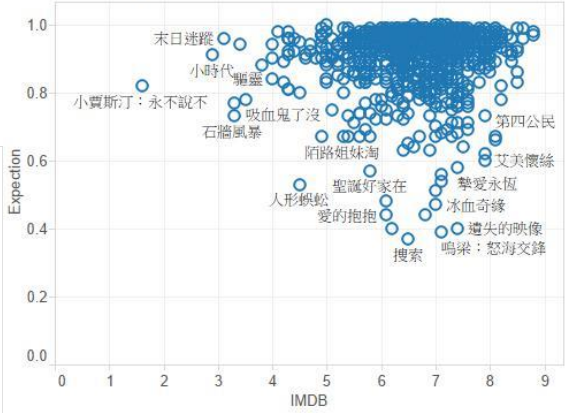




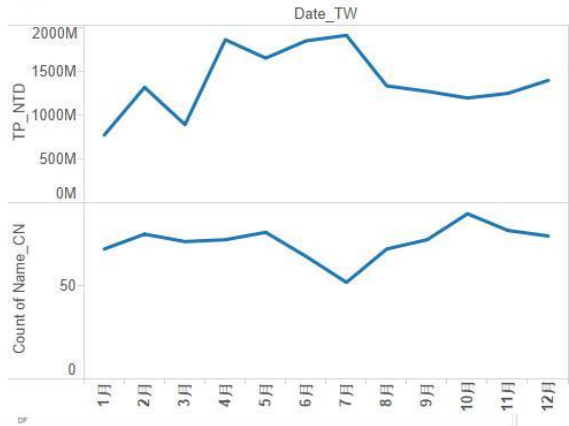
Distinct count of TP_NTD for each Country.

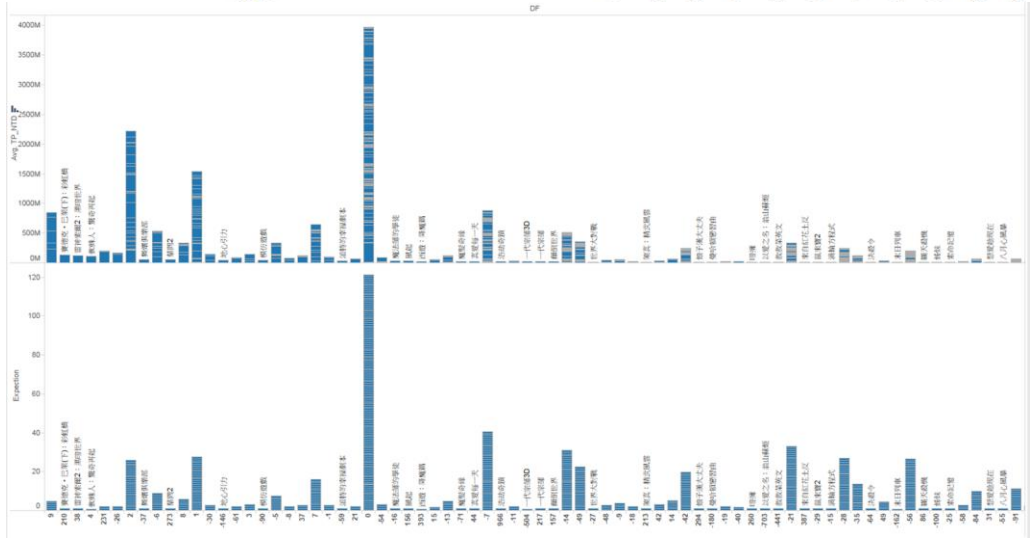


IMDB v.s. expectation



by Month



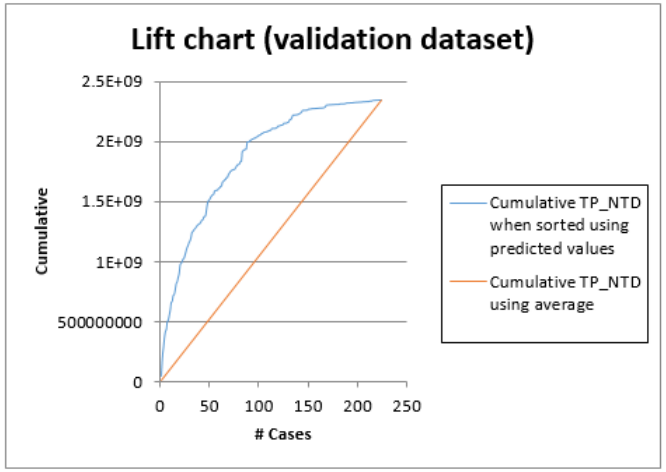


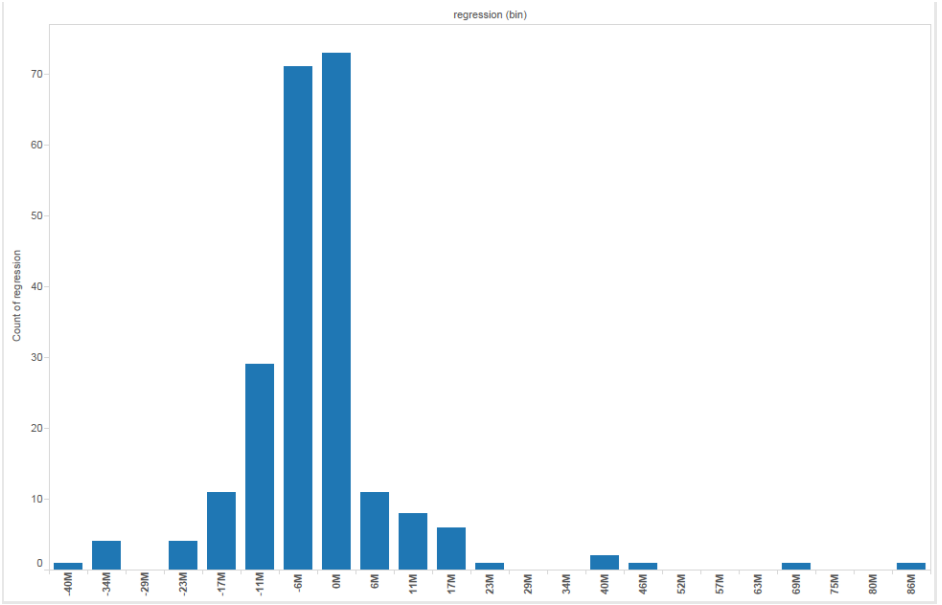
Method

Comparing performance of validation dataset

Prediction method:

1. Linear regression
2. Neural Network
3. Random Trees





Subset 1
Training v.s. validation 60:40

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
4.30769E+16	11322761.65	-3.68233E-09

Validation Data Scoring - Summary Report

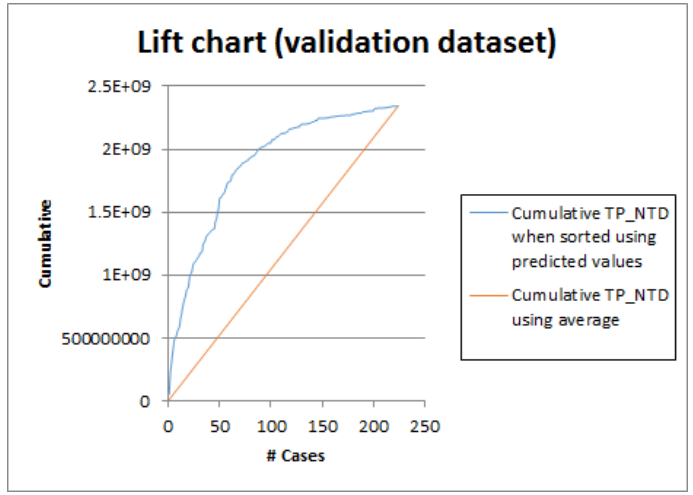
Total sum of squared errors	RMS Error	Average Error
3.58447E+16	12649946.6	-22719.81588

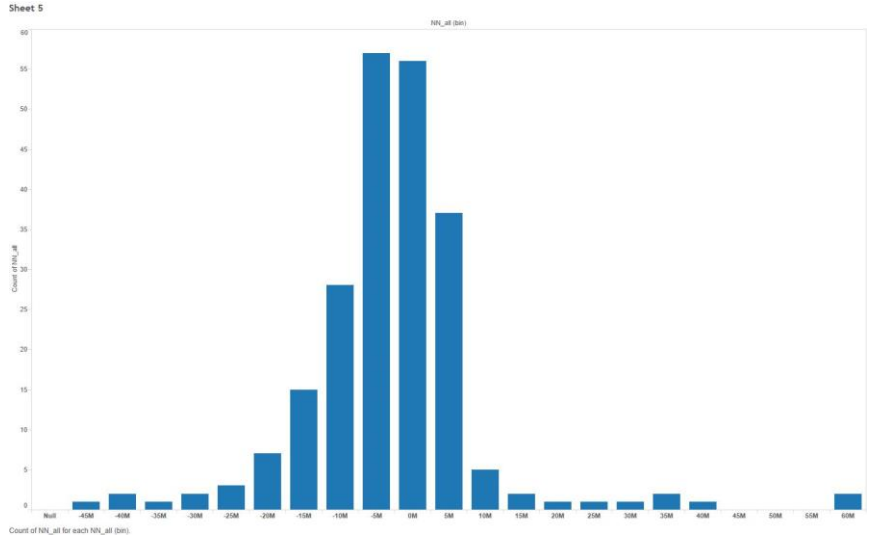
Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	-12697073.9	4673393.378	-2.71689	0.00694	-21890679.99	-3503467.728	3.44087E+16
Budget_USD	0.106243247	0.024346869	4.363733	1.72E-05	0.05834753	0.154138964	3.06376E+16
USA_USD	0.120776433	0.014954132	8.076459	1.28E-14	0.091358324	0.150194543	1.99474E+16
IMDB	1595308.528	687850.0483	2.319268	0.020994	242154.2114	2948462.844	3.21788E+13
YouTube Trailor PV	40.35468794	3.831109006	10.53342	1.53E-22	32.81804285	47.89133302	1.77227E+16
Caring	-8419909.15	3927120.661	-2.14404	0.032764	-16145430.56	-694387.7403	6.53148E+14
Action	6586576.703	1856439.072	3.547963	0.000445	2934547.383	10238606.02	2.6168E+15
Animation	-10718187.6	2957450.373	-3.62413	0.000336	-16536151.46	-4900223.694	1.72496E+15

Comparing NN and random trees

Validation Data scoring - Summary Report

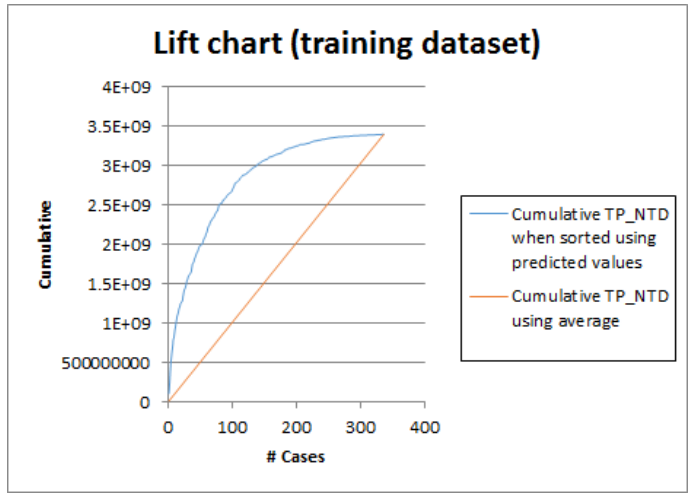
Total sum of squared errors	RMS Error	Average Error
3.32205E+16	12178078.6	-411749

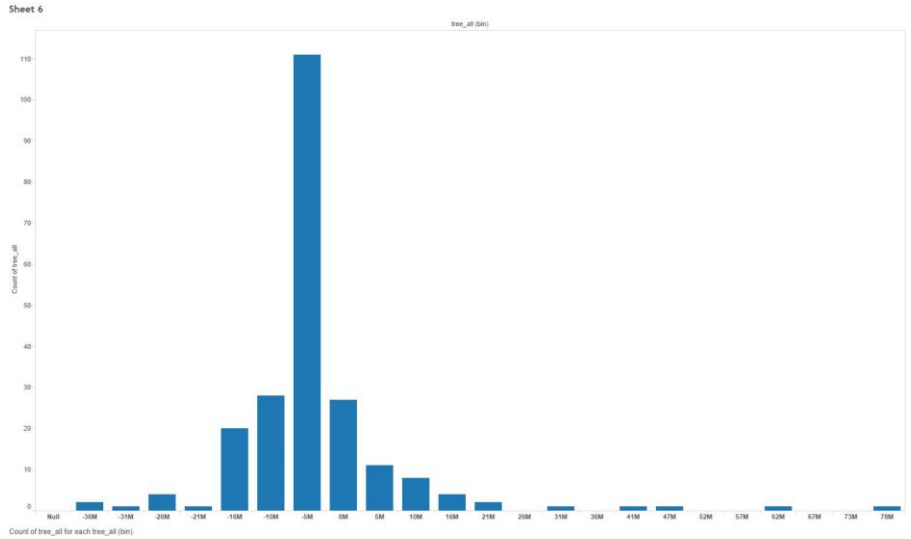




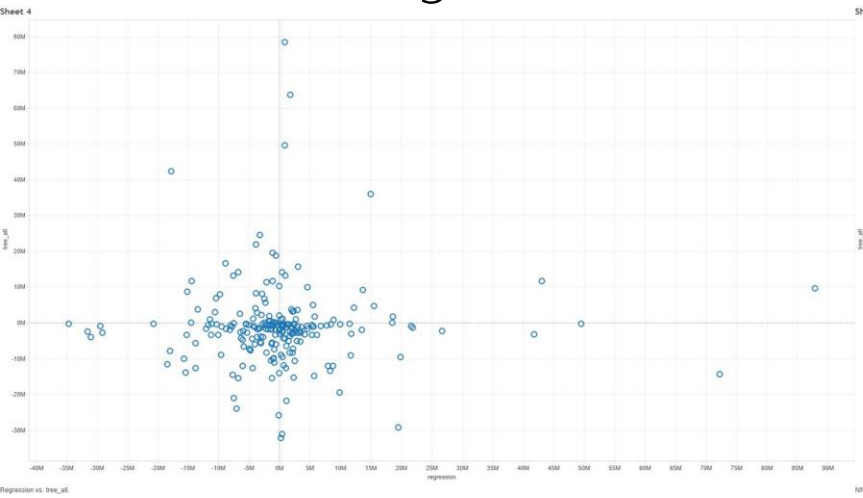
Validation Data scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
3.13972E+16	11839168.7	-1036429.5

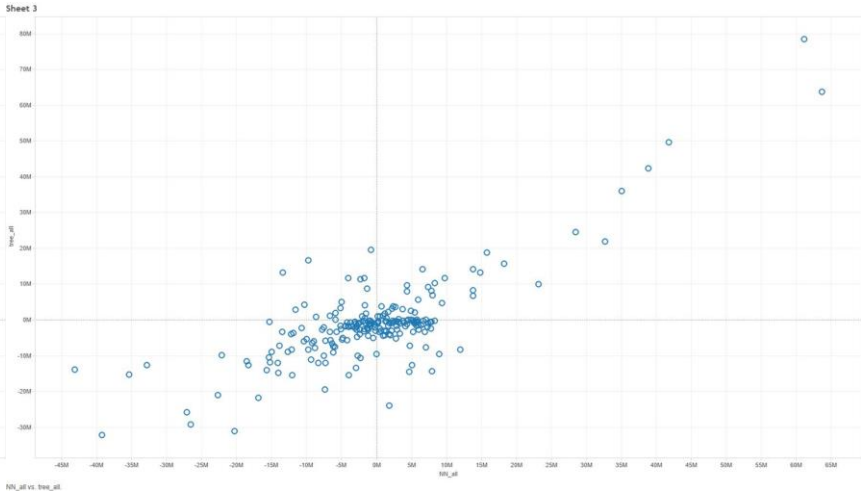




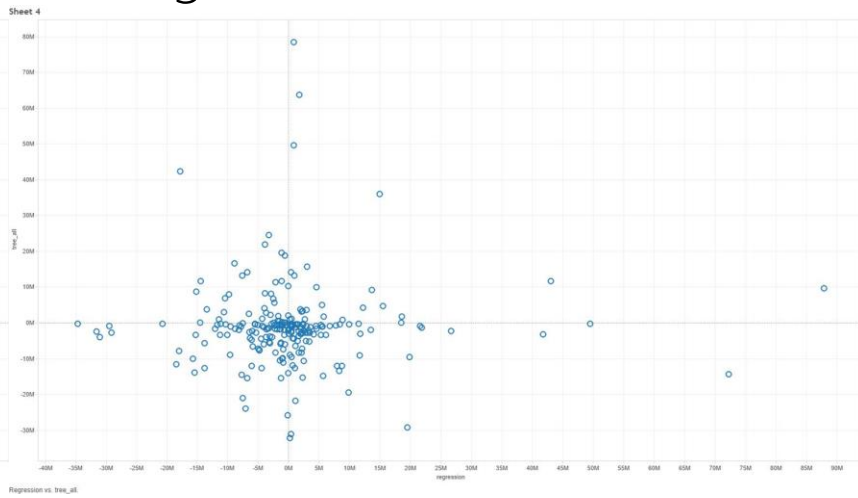
NN vs. regression



NN vs. random trees



regression vs. random trees



Ensemble

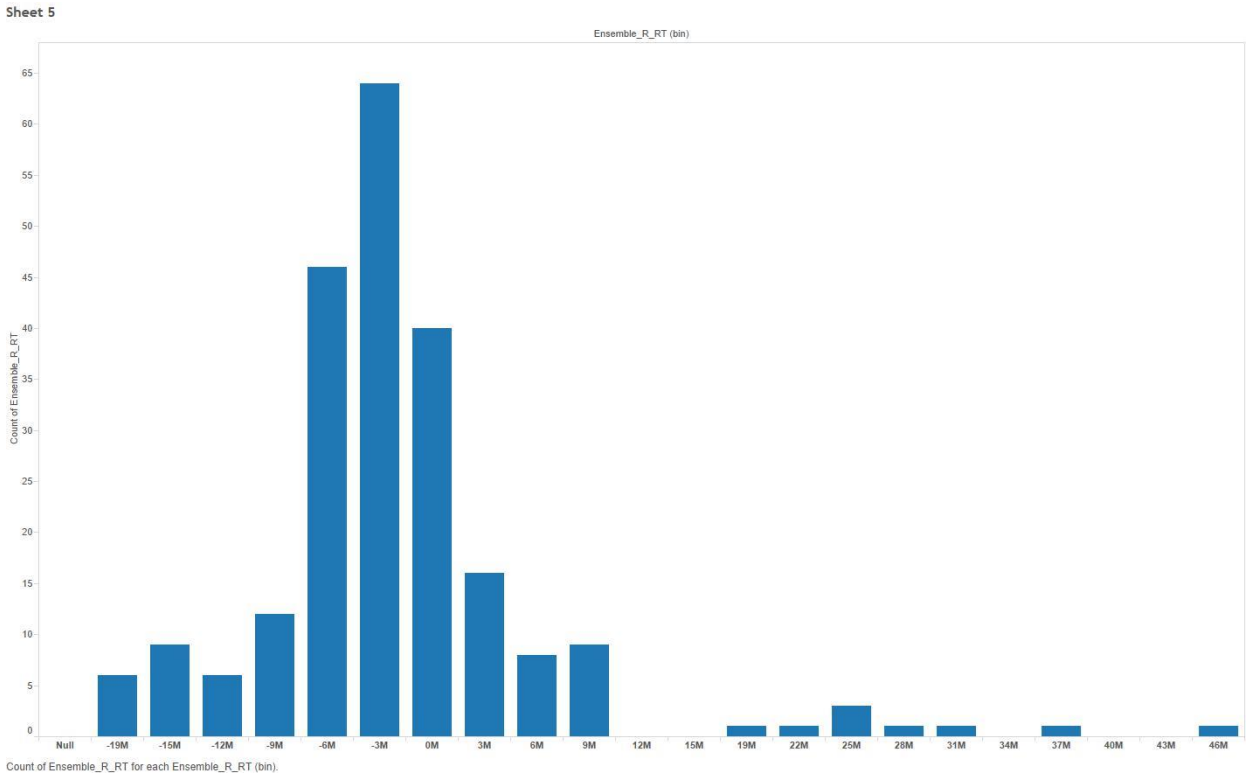
We ensemble regression and random trees. And we get the better performance.

Performance evaluation

Evaluation method:

1. Residual histogram
2. RMSE / Average Error

Total sum of squared errors	RMS Error	Average Error
1.6869E+16	8678017.449	-529574.6658



Residual histogram

New data prediction

Name_CN	Name_EN	Regression	RT	Ensemble	Current	Date_TW	Length	Director	Cast	Agent	Expection	Type
史努比	A Peanuts Movie	14,663,134	41,219,598	27,941,366	8,510,000	12/24/2015	88	《冰原歷險記4：	福斯影片	94%	動畫
紐約愛未眠	Before We Go	4,494,963	5,210,720	4,852,842	890,000	12/24/2015	89	克里斯伊	《美國隊	采昌國際	94%	劇情/喜
真相急先鋒	Truth	2,390,408	11,544,750	6,967,579	2,730,000	12/24/2015	125	詹姆斯范	《藍色茉	傳影互動	100%	劇情
翻轉幸福	Joy	46,223,196	48,491,210	47,357,203	1,210,000	12/31/2015	124	《派特的	《飢餓遊	福斯	96%	劇情
家有兩個爸	DADDY'S HOME	11,804,761	12,069,606	11,937,184		12/31/2015	96	《老闆不	《官賤對	派拉蒙影	91%	喜劇
怪物遊戲	Goosebumps	11,462,122	11,336,878	11,399,500		12/31/2015	103	《鯊魚黑	《格列佛	索尼影業	86%	奇幻/劇
神鬼獵人	The Revenant	17,626,759	35,740,316	26,683,538		1/8/2016	151	阿利安卓	李奧納多	CatchPlay	92%	劇情
瞎趴姊妹	Sisters	9,839,438	14,586,851	12,213,144		1/8/2016	118	《歌喉讚	《愛在頭	環球影業	100%	喜劇
女權之聲：	Suffragette	3,466,417	11,632,180	7,549,299		1/8/2016	106	莎拉賈芙	《大亨小	絕色國際	100%	劇情
45年	45 Years	(320,258)	3,627,123	1,653,432		1/15/2016	95	《愛在週	《里斯本	傳影互動	89%	劇情
大賣空	The Big Short	7,444,493	13,886,539	10,665,516		1/15/2016	130	《銀幕大	《黑暗騎	派拉蒙影	100%	劇情
史帝夫賈伯	Steve Jobs	9,829,875	6,676,163	8,253,019		1/22/2016	122	《貧民百	《X戰警	環球影業	86%	劇情
鼠來寶：鼠	Alvin and The Chi	10,128,811	20,359,284	15,244,047		1/22/2016	86	《荒野大	傑森李/	福斯影業	95%	喜劇
恐龍當家	The Good Dinosau	22,078,416	35,633,186	28,855,801		2/5/2016	93	Peter Sol	(配音)茱	迪士尼	93%	冒險/喜
扣押幸福	Freeheld	(1,981,177)	3,697,544	858,183		2/19/2016	104	《愛情無	《我想念	車庫娛樂	96%	劇情/愛
驚爆焦點	Spotlight	5,862,277	13,451,823	9,657,050		2/19/2016	128	《幸福來	《鳥人》	采昌國際	100%	劇情
八惡人	The Hateful Eight	5,967,339	18,697,211	12,332,275		2/19/2016	182	昆汀塔倫	山繆傑克	CatchPlay	100%	劇情/懸
丹麥女孩	The Danish Girl	11,524,626	17,207,467	14,366,046		3/4/2016	120	《王者之	《愛的基	環球影業	98%	劇情

Recommendations

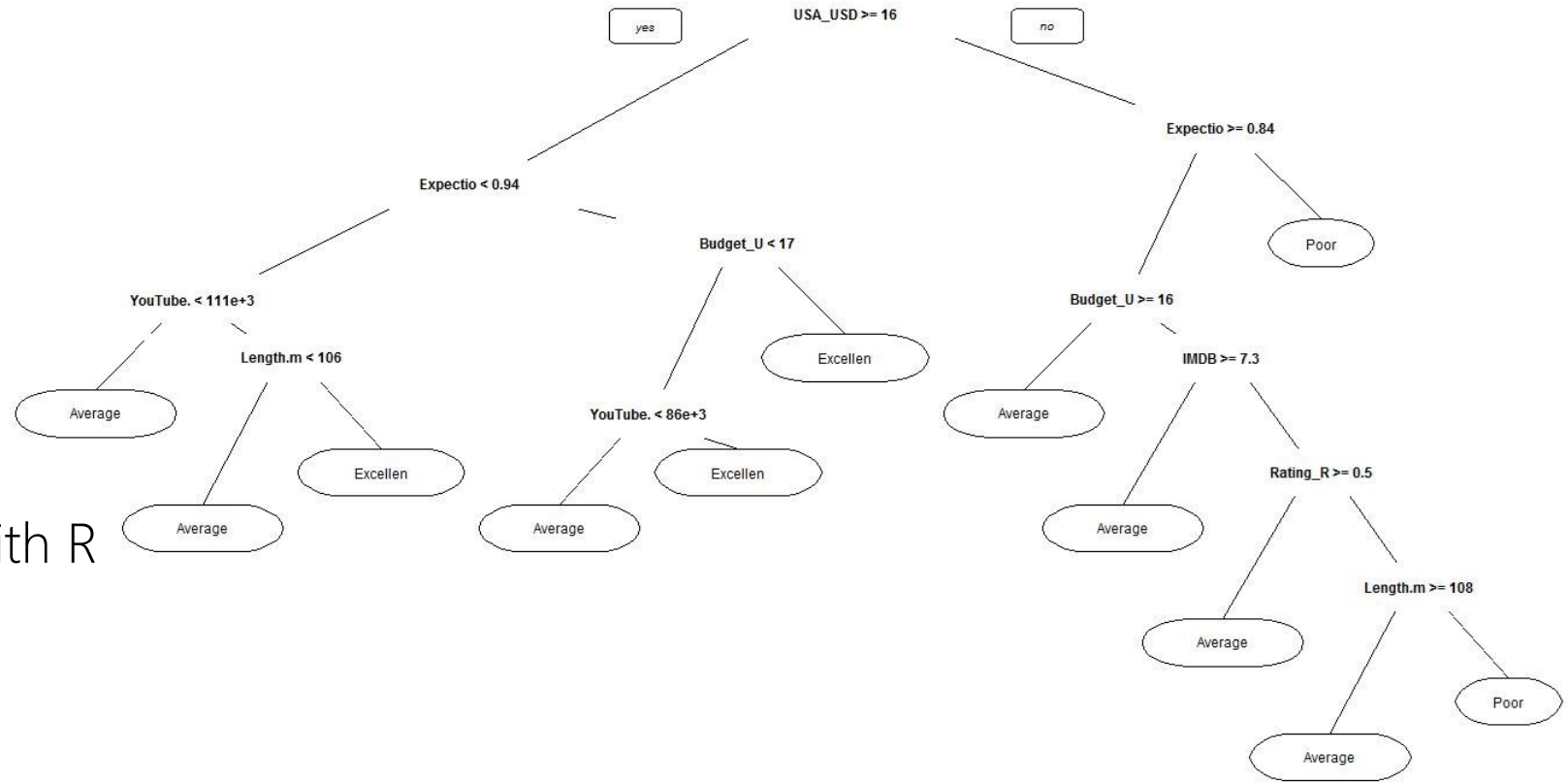
1. Should collect more data…
2. Lots of missing values in different dimensions
3. Released weeks of movies should be considered but are hard to collect
4. Expectation and IMDB should concern about number of rating consumers
5. Youtube trailer pageviews are accumulated through years
6. Floating exchange rate
7. Increase of movie ticket price and change of consumer behavior
8. Need to seek for more domain knowledge to get predictors with valuable insights
9. Could try to change this task as a classification problem

```
           Average Excellent Poor
  Average       55        10   16
  Excellent     12        32    0
  Poor          17         0   26
> print(paste('Accuracy on the validation set', (55+32+26)/nrow(validation)))
[1] "Accuracy on the validation set 0.672619047619048"
>
> table(validation$Performance)

  Average Excellent      Poor
       84        42        42
> print(paste('Baseline model accuracy on the validation set', (84)/nrow(validation)))
[1] "Baseline model accuracy on the validation set 0.5"
```

Classification tree

We also tried classification tree with R



Thank you for your attention!