Prediction of the movie revenue

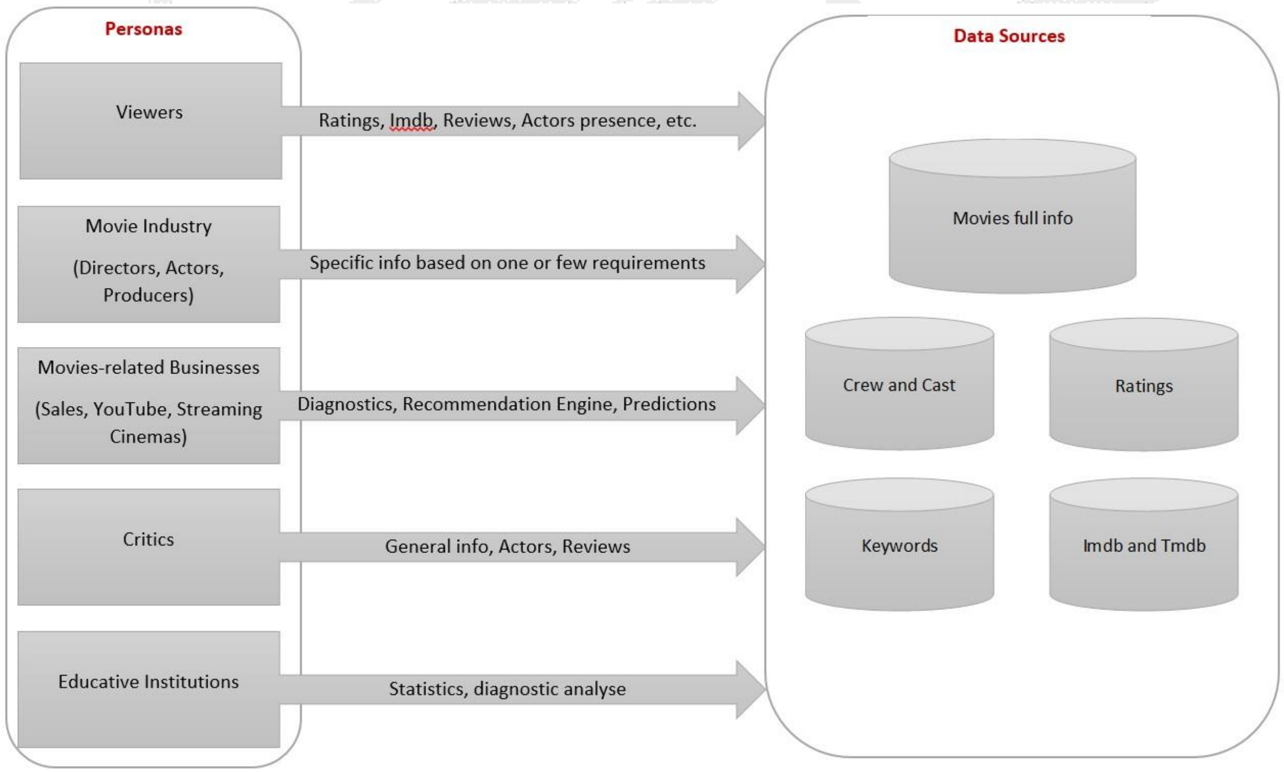
1 PROBLEM DEFINITION

1.1 INTRODUCTION

- Big data to improve greenlighting, budgeting and marketing

1.2 AREAS OF USAGE OR BUSINESS USE-CASES

- Make educated guesses (ticket sales, profit margins, reviews, social chatter, franchise options, awards ...)



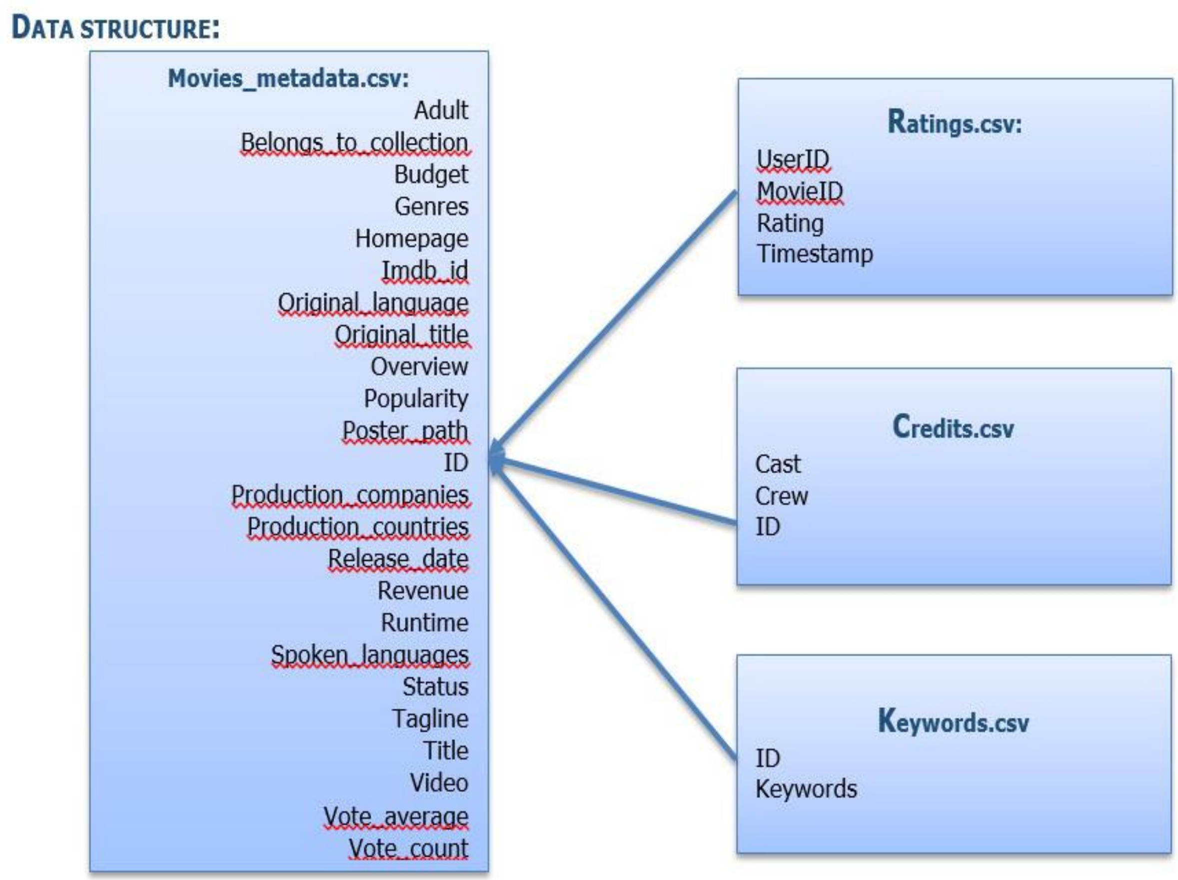
2 DATASET DESCRIPTION

2.1 SOURCE

- https://www.kaggle.com/rounakbanik/the-movies-dataset

2.2 FILES AND STRUCTURE

- 45 000 movies metadata from the Full MovieLens Dataset up to July 2017
- 26 million ratings from 270 000 users from the GroupLens website



3 APPROACH

The Movies Dataset gives us an opportunity to train on data preprocessing, perform statistical analyse and discover new machine learning methods. Its structure is somehow complicated and doesn't always follow logical direction. On another hand, these challenges helped us improve our skills.

Here are the features of the dataset that we will be using:

	id	title	budget	revenue	production_countries	release_date	popularity	vote_average	vote_count	genres	production_companies	belongs_to_collection	cast	keywords	year	year_month
3	121173	Voracious	11178	34659.000	[{'iso_3166_1': 'PH', 'name': 'Philippines'}]	2012-09-05	0.079	8.000	1.000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	[{'name': 'APT Entertainment', 'id': 8355}, {'...	0	[{'cast_id': 16, 'character': 'Rene', 'credit_...	[{'id': 4694, 'name': 'staged death'}, {'id': ...	2012	2012-09
5	110428	Camille Claudel 1915	3512454	115860.000	[{'iso_3166_1': 'FR', 'name': 'France'}]	2013-03-13	0.110	7.000	20.000	[{'id': 18, 'name': 'Drama'}]	[{'name': 'Canal+', 'id': 5358}, {'name': 'Art...	0	[{'cast_id': 3, 'character': 'Camille Claudel'...	[{'id': 254, 'name': 'france'}, {'id': 745, 'n...	2013	2013-03
6	110428	Camille Claudel 1915	3512454	115860.000	[{'iso_3166_1': 'FR', 'name': 'France'}]	2013-03-13	0.110	7.000	20.000	[{'id': 18, 'name': 'Drama'}]	[{'name': 'Canal+', 'id': 5358}, {'name': 'Art...	0	[{'cast_id': 3, 'character': 'Camille Claudel'...	[{'id': 254, 'name': 'france'}, {'id': 745, 'n...	2013	2013-03

Example of ***cast*** feature for one movie (one cell of the database):

[{'cast_id': 1, 'character': 'Felix the Cat (Voice)', 'credit_id': '52fe45f09251416c91043a2f', 'gender': 0, 'id': 115502, 'name': 'Chris Phillips', 'order': 0, 'profile_path': '/67G2MHc1Qs2ox2PDyQFxEhVSgYp.jpg'}, {'cast_id': 2, 'character': 'Princess Oriana (Voice)', 'credit_id': '52fe45f09251416c91043a33', 'gender': 0, 'id': 115503, 'name': "Maureen O'Connell", 'order': 1, 'profile_path': None}, {'cast_id': 3, 'character': 'The Duke of Zill / Wack Lizardi (voice) (as Peter Neuman)', 'credit_id': '52fe45f09251416c91043a37', 'gender': 0, 'id': 115504, 'name': 'Peter Newman', 'order': 2, 'profile_path': None}, {'cast_id': 5, 'character': '(Voice)', 'credit_id': '52fe45f09251416c91043a3b', 'gender': 0, 'id': 115506, 'name': 'Susan Montanaro', 'order': 4, 'profile_path': None}, {'cast_id': 6, 'character': '(Voice)', 'credit_id': '52fe45f09251416c91043a3f', 'gender': 0, 'id': 115507, 'name': 'Don Oriolo', 'order': 5, 'profile_path': None}, {'cast_id': 7, 'character': '(Voice)', 'credit_id': '52fe45f09251416c91043a43', 'gender': 0, 'id': 48402, 'name': 'Christian Schneider', 'order': 6, 'profile_path': None}, {'cast_id': 8, 'character': '(Voice)', 'credit_id': '52fe45f09251416c91043a47', 'gender': 0, 'id': 115508, 'name': 'David Kolin', 'order': 7, 'profile_path': None}, {'cast_id': 9, 'character': '(Voice)', 'credit_id': '52fe45f09251416c91043a4b', 'gender': 0, 'id': 115509, 'name': 'Michael Fremer', 'order': 8, 'profile_path': None}, {'cast_id': 10, 'character': 'Madam Pearl (voice) (as Alice Playton)', 'credit_id': '52fe45f09251416c91043a4f', 'gender': 1, 'id': 80165, 'name': 'Alice Playten', 'order': 9, 'profile_path': '/oRaMq0i9PI64VKPVivVe0xCPDKB.jpg'}]

173 words

As you can see extracting information from columns is time consuming.

3.1 Data cleaning

What we did:

- Removed NANs
- Removed 0 and small revenues and budget
- Replaced outliers (checked numbers on IMDB)
- Kept only *released* movies (*removed rumored, post-production*)
- Removed movies without specified production companies
- Merged the data

What we discovered:

- Issue with currency (budget is in different currencies). Removed all the movies were USA wasn't in production countries
- Overestimation of popularity: this feature fluctuates with time. So we cannot rely on it for predictions.

3.2 Feature Engineering

Here are some main points in our process of feature engineering:

- Budget, year (numeric)
- Dummies (genres, production companies, actors, belongs_to_collection)
- Applied log to the *budget*
- Created new numeric features:
    - Collection votes
    - Genres average vote, average vote count and average revenue
    - Production companies average vote, average vote count and average revenue
    - Actors average vote, average vote count and average revenue

Here is the pattern that we used:

Original dataset

Movies	Genres	Revenue	Average vote	Vote count
Movie-1	Action, SF, Drama	100	6.7	1000
Movie-2	Action, Comedy	200	7.3	800
Movie-2	SF, Drama	300	7.5	2000

Numeric representations of genres data

Export to API

Genre_name	Average Revenue	Average Ave_vote	Average Vote_count
Action	(100+200)/2 = 150	(6.7+7.3)/2 = 7	(1000+800)/2 = 900
Comedy	200	7.3	800
SF	200	7.1	1500
Drama	200	7.1	1500

Features we can use for prediction

	Action	Comedy	SF	Drama	Average Revenue	Average Ave_vote	Average Vote_count
Movie-1	1	0	1	1	(150+200+200)/3=183.3	(7+7.1+7.1)/3=7.07	(900+1500+1500)/3=1300
Movie-2	1	1	0	0	125	7.15	850
Movie-3	0	0	1	1	200	14.2	1500

Instead of having only one valuable numeric value we have now 11. It improved our results drastically.

4 Prediction and Results

Finally we have a model with 254 features. For prediction we used RandomForests regressor with n_estimators=500.

4.1 Final results

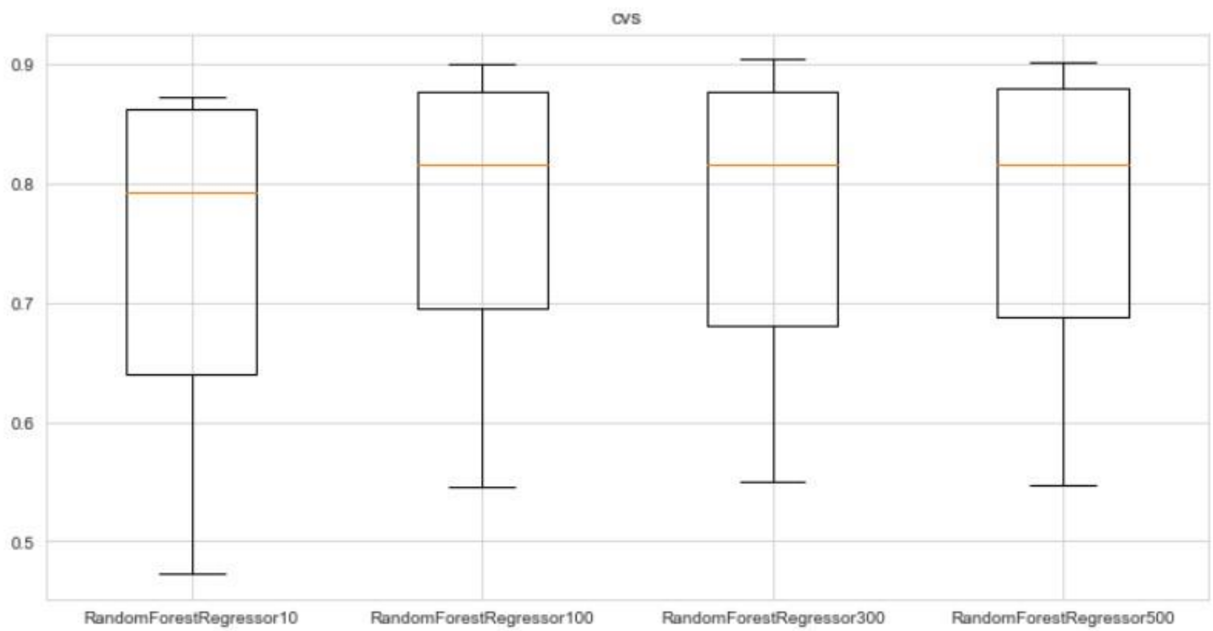
CVS

```
scores = cross_val_score(RandomForestRegressor(500), X, y, cv=10)
print('cross_val_score', np.mean(scores))
```

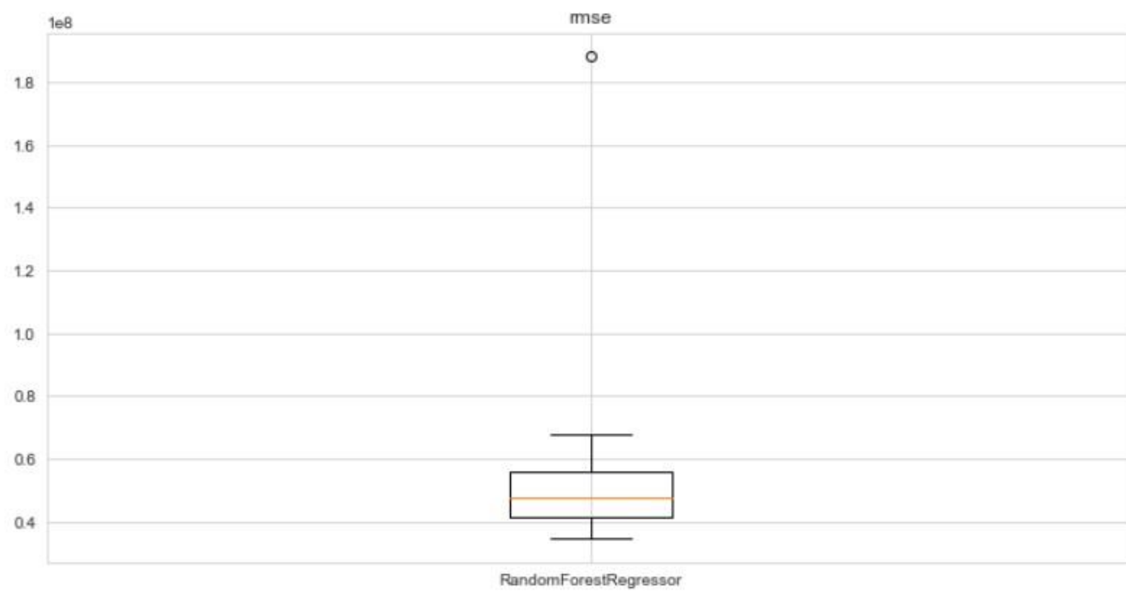
```
cross_val_score 0.7775488513694426
```

Boxplots of CVS:

```
MODEL RandomForestRegressor10
MODEL RandomForestRegressor100
MODEL RandomForestRegressor300
MODEL RandomForestRegressor500
```

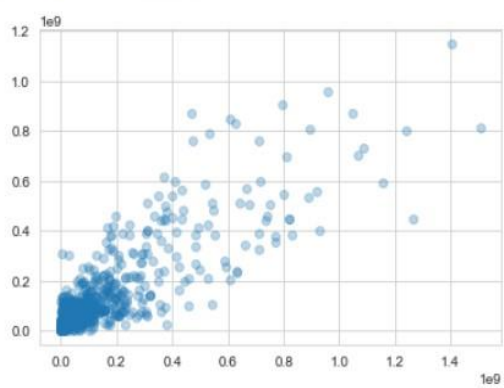


RMSE



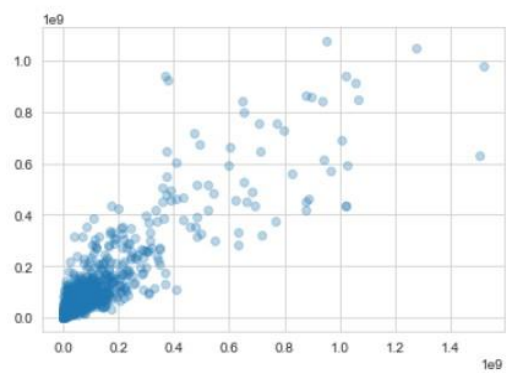
Scatter plot of real values VS predicted at different stages of feature engineering:

MAE 63336750.16157023



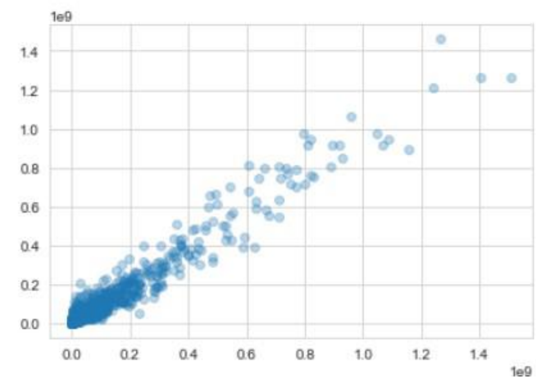
Budget and dummies

MAE 54387893.99516461



Budget, dummies and numeric values except actors

MAE 31605379.642912365



All plus actors

4.2 Feature importance and error correlation

	0	1
253	a_rev_ave	0.893
128	p_rev_ave	0.023
3	coll_vote	0.012
0	budget	0.011
251	a_vote_count	0.009
23	g_vote_count	0.006
252	a_vote_ave	0.004
126	p_vote_count	0.004
1	year	0.003
24	g_vote_ave	0.003
127	p_vote_ave	0.003
25	g_rev_ave	0.003
77	Ingenious Film Partners	0.002

```
Animation                0.210
Adventure                0.216
collection               0.231
g_vote_count             0.239
g_rev_ave                0.292
coll_vote                0.357
budget                   0.383
p_vote_count             0.393
a_vote_count             0.422
p_rev_ave                0.491
revenue                  0.575
prediction               0.577
a_rev_ave                0.579
abs_error                1.000
Name: abs_error, Length: 256, dtype: float64
```

4.3 Real movies testing using application

Predicted Revenue

COMPANY: Warner Bros.

COLLECTION: Ocean's Collection

YEAR: 2018

BUDGET: 77,000,000.0

GENRES: Action, Adventure, Thriller

ACTORS: Cate Blanchett, Anne Hathaway

369,336,550.3

PREDICTED REVENUE

Ocean's Eight Real Revenue: $297,718,711

Predicted Revenue

COMPANY:

COLLECTION: Hotel Transylvania Collection

YEAR: 2018

BUDGET: 80,000,000.0

GENRES: Animation, Adventure, Comedy

ACTORS: Adam Sandler, Steve Buscemi

518,974,493.8

PREDICTED REVENUE

about 538,000,000 Revenue

4.4 Possible improvements

- Ratings, drawback: many NANs
- keywords, drawback: many NANs
- External data: cast credits