Ecole polytechnique fédérale de Lausanne

Data Science for Business: predict IPO shares prices

Project presentation - Team 11

Guillaume Grandjean, Julien Belguise, Mariem Belhaj Ali

Table of contents

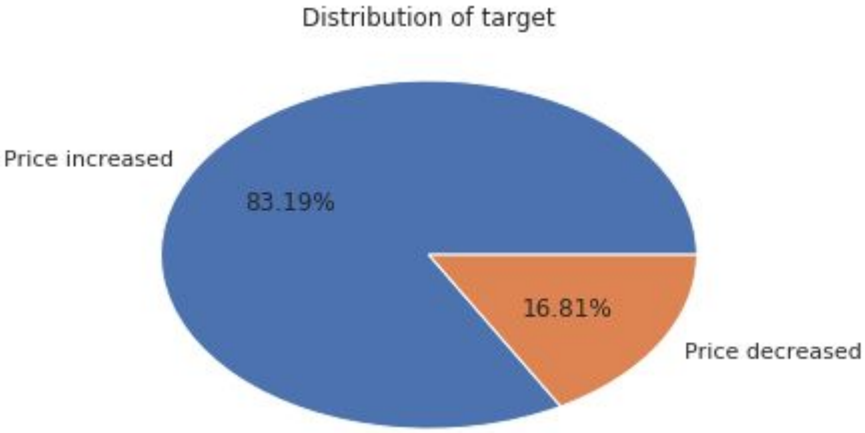
- Data: datasets, pre-procesing, missing values, text features
- Classification with non-textual features
- Classification with textual features
- Classification with all features
- Price prediction
- How it help us to invest ?

Data: datasets, pre-processing, missing values

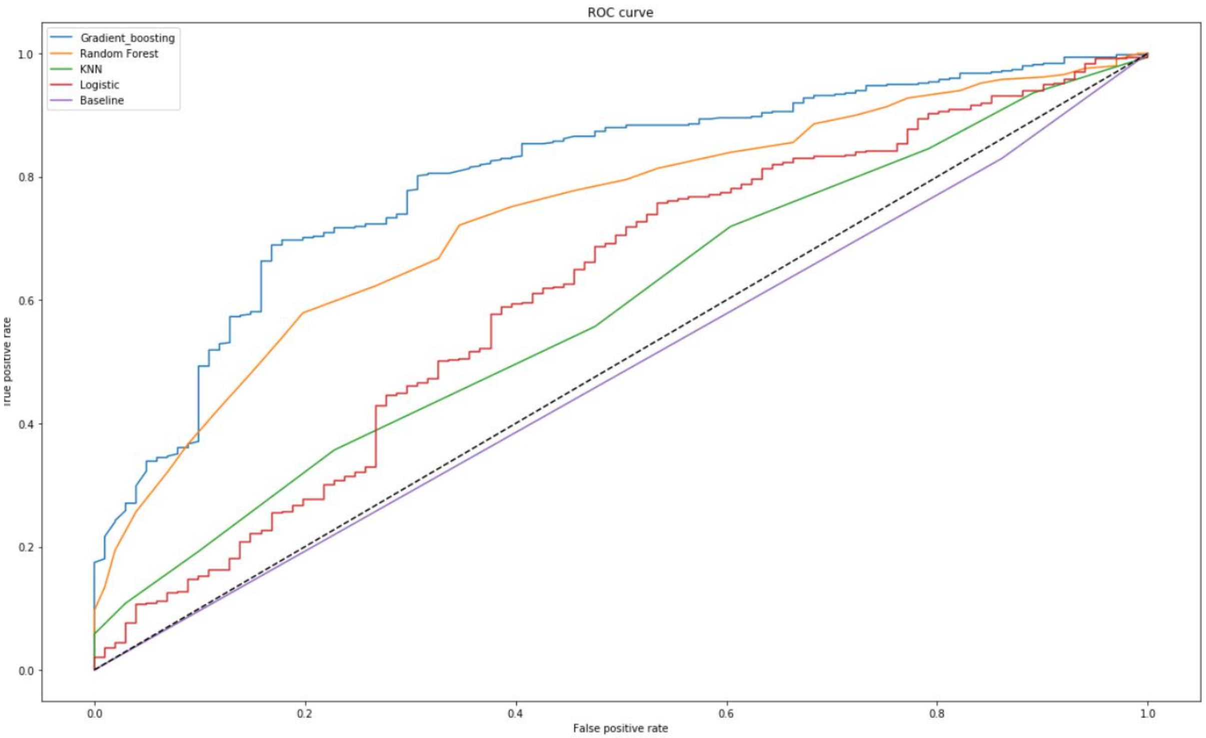
Beginning shape of dataset
3000x**159**

ROC AUC to score
target unbalanced

- ➢**Add dataset**
  - -> Zip code for every us state
  - -> Categorization of the types of Industry
  - -> Interest rate of us bond
- ➢ **Handle Missing Values**
  - -> Drop features with more than 50% of missing values
  - -> Use most frequent Value to fill the rest
- ➢ **Feature reduction**
  - -> Correlation



Classification with non-textual features



Roc Auc Scores:

- **0.847 Gradient Boosting**
- 0.737 Random Forest
- 0.643 Logistic
- 0.584 KNN
- 0.484 Baseline

Outcome:
Pretty good score with Gradient Boosting

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Classification with textual feature: Risk factors

Analyzing relevant information about risk factors that might affect future business performances to predict price increase:

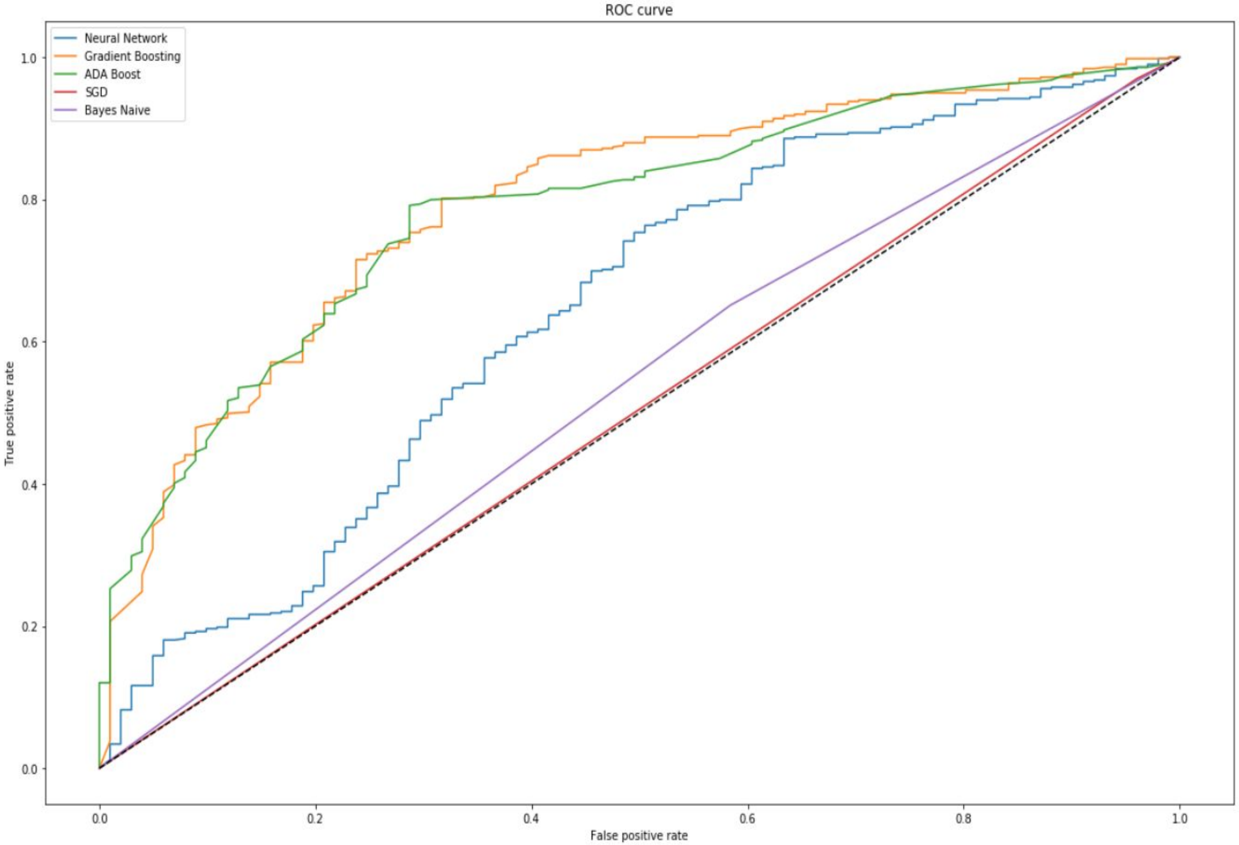
- Cleaning the text
- Numerizing it with a TFIDF to reflect how important a word is
- Training our models

Roc Auc Scores:

- 0.631 XGB
- 0.5911 Random Forest
- 0.643 Random Forest with PCA
- 0.584 XGB with PCA

Outcome:
Scored less with textual features than with non textual features

Classification with all features



Roc Auc Scores:

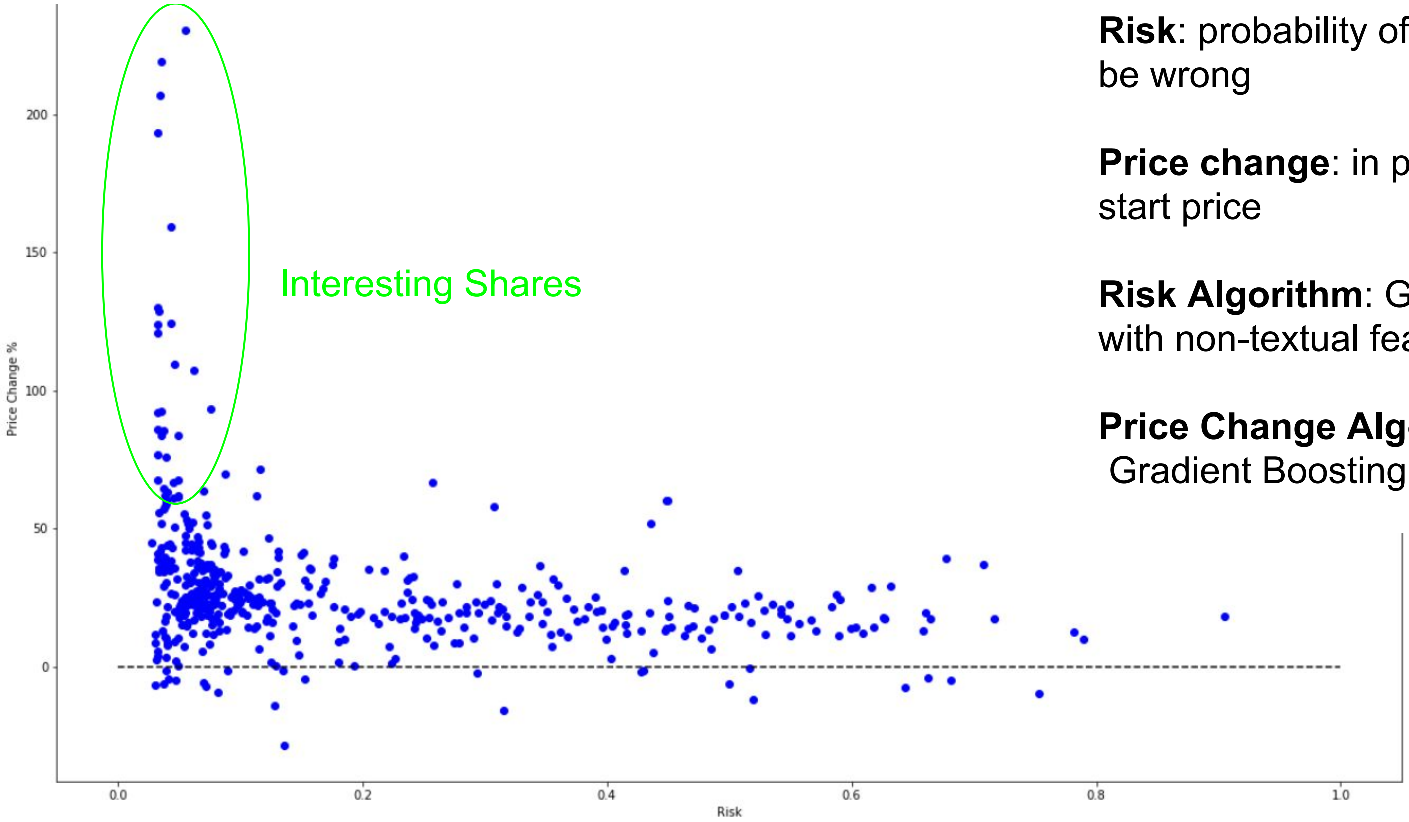
- 0.796 Ensemble ADA+GB
- 0.793 Gradient Boosting
- 0.785 ADA
- 0.648 Neural Network
- 0.534 Bayes
- 0.505 SGD

Outcome:
Scored less with non textual features than with only textual features.

Price prediction

Best Model	Gradient Boosting Regressor
R-squared score	0.88
Mean Square Error	0.0023
Mean Absolute Error	0.032
Median Absolute Error	0.021

How it help us to invest ?



Risk: probability of our prediction to be wrong

Price change: in percentage of the start price

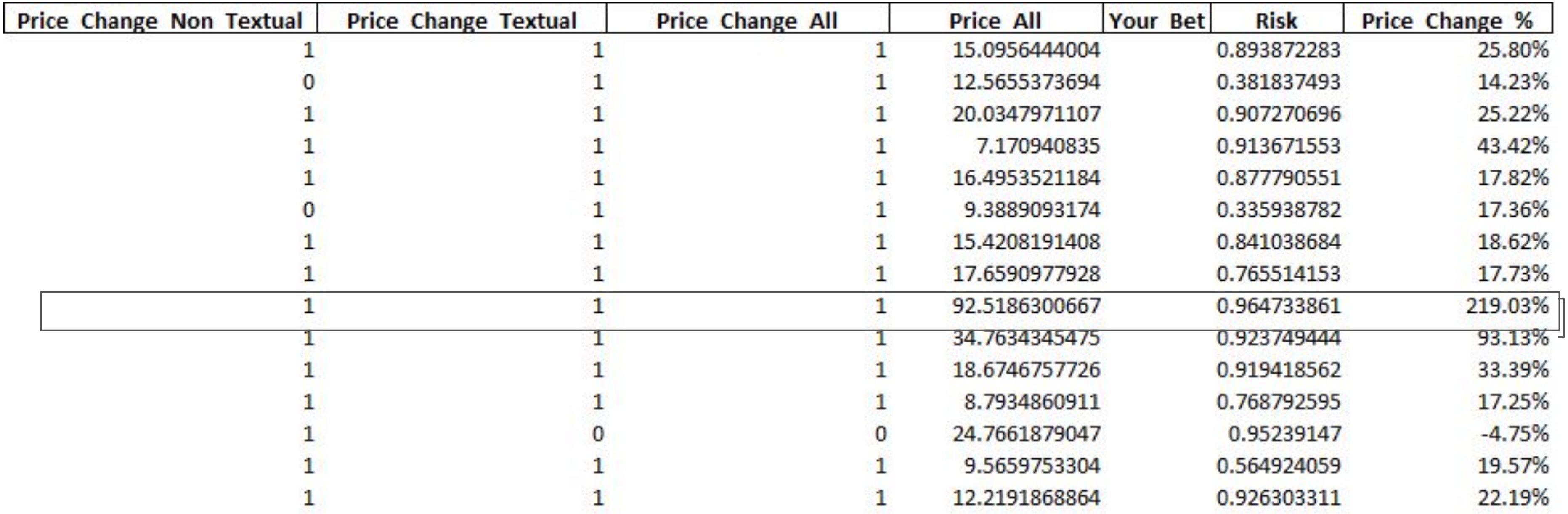
Risk Algorithm: Gradient Boosting with non-textual features

Price Change Algorithm:
Gradient Boosting regressor

How it help us to invest ?

Depending on your Investment Strategy :

➜ Select **Highest** Price Change with **Lowest** Risk



Price Change Non Textual	Price Change Textual	Price Change All	Price All	Your Bet	Risk	Price Change %
1	1	1	15.0956444004		0.893872283	25.80%
0	1	1	12.5655373694		0.381837493	14.23%
1	1	1	20.0347971107		0.907270696	25.22%
1	1	1	7.170940835		0.913671553	43.42%
1	1	1	16.4953521184		0.877790551	17.82%
0	1	1	9.3889093174		0.335938782	17.36%
1	1	1	15.4208191408		0.841038684	18.62%
1	1	1	17.6590977928		0.765514153	17.73%
1	1	1	92.5186300667		0.964733861	219.03%
1	1	1	34.7634345475		0.923749444	93.13%
1	1	1	18.6746757726		0.919418562	33.39%
1	1	1	8.7934860911		0.768792595	17.25%
1	0	0	24.7661879047		0.95239147	-4.75%
1	1	1	9.5659753304		0.564924059	19.57%
1	1	1	12.2191868864		0.926303311	22.19%