IPO Underpricing Prediction

Data Science for Business

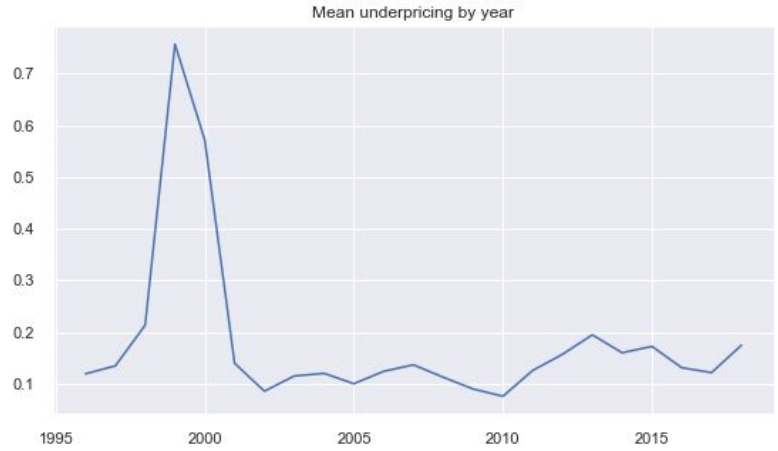
Team : Paul Mermod
Smail Ait Bouhsain
Tomas Giro Larraz

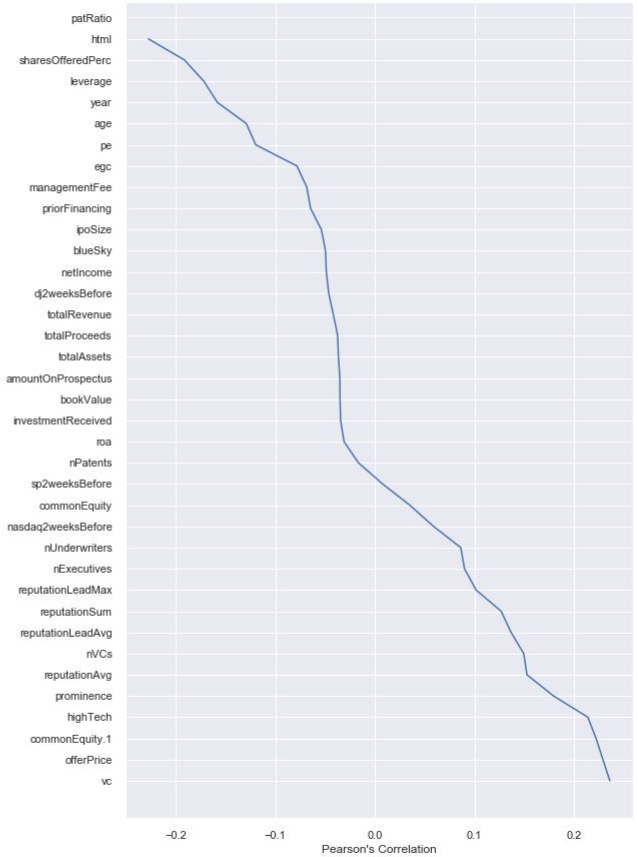
Table of contents

Data visualization

*Goal: Identify **correlations** between features and target*

- Target: % Change in stock price on first day
- Motivation: inclusion of these features in the models



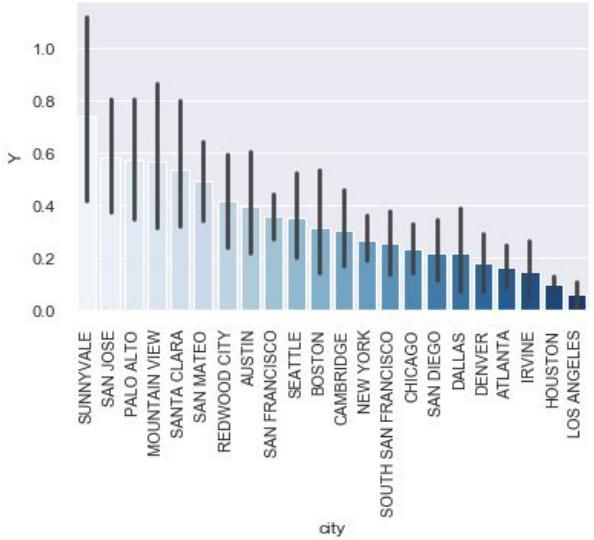


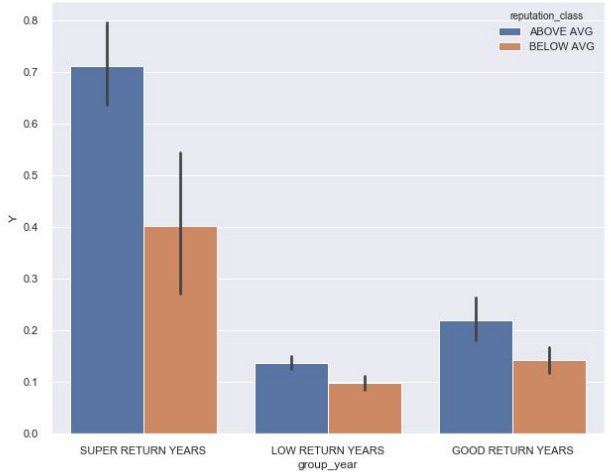
Correlation with Target on “data-to-train”

Feature extraction

Motivation: captures predictive information by hand-engineering features

- ➔ **Categorical**:
  - Create dummies for different groups on their average first day return: e.g. super_return>years, good_return>years, city>super_return, city>good_return, industryFF12, exchange market.
  - Model interaction between different features: Effect of reputation on target not constant across years.
- ➔ **Skewed continuous**: Add the logarithm to the dataset to model non-linear relationship
- ➔ **Remove redundant and correlated attributes**
- ➔ **Deal with missing data using Linear Regression**

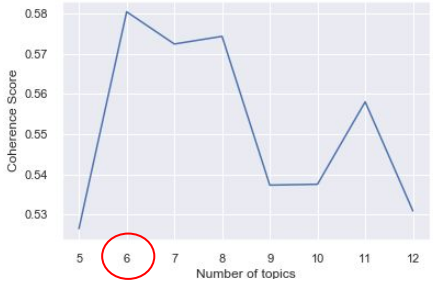


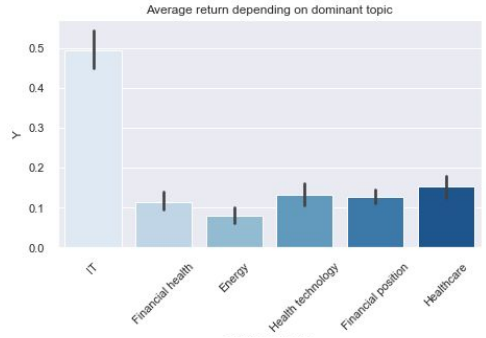


Feature Extraction: modelling risk factors

Motivation: extract predictive information from risk factors through NLP techniques.

- **LDA Model**: Assume that each latent topic refers to a specific risk factor. Optimizing the number of topics based on coherence score.
  - Look at dominant topic of each risk factor document
  - Risk factors whose dominant topic is related to “IT” have a significant higher average return (i.e. overpricing)

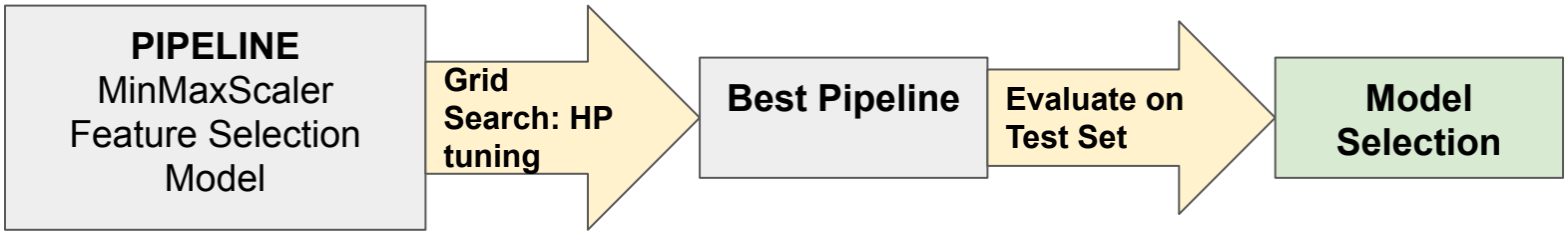




- Engineer variables reflecting risk: based on utilization of certain words relative to rf text length.
  - Indebtedness, patRatio, reimbursement, default, trial

ML system

- **Baseline models**:
  - Classification tasks: ***Dummy Classifier, K-Nearest Neighbors Classifier***
  - Regression task: ***Linear Regression***
- **Advanced models**:
  - Classification tasks: ***Logistic Regression Ridge & Lasso, Random Forest Classifier***
  - Regression task: ***Ridge, Lasso, Support Vectors Regressor, Random Forest Classifier***
    - Feature selection: **SelectKBest(f_classif) for Random Forest, L1/L2 penalty otherwise**



Performance metrics & Model selection

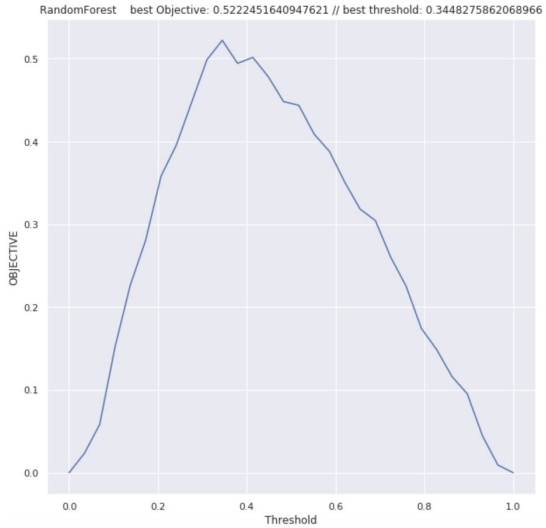
Performance metrics:

- Deterministic Classifiers P1,..,P5: **roc AUC score**
- Regression P6: **Mean Squared Error**
- Probabilistic Classifiers P7,..,P9: **custom metrics** based on deviation of probability

For P1,...,P5: *Optimize Probability threshold for the best classifier:*

Optimal Threshold = Argmax(TPR-FPR)

- This score maximizes “Hit Rate” while minimizing “False alarm Rate”.



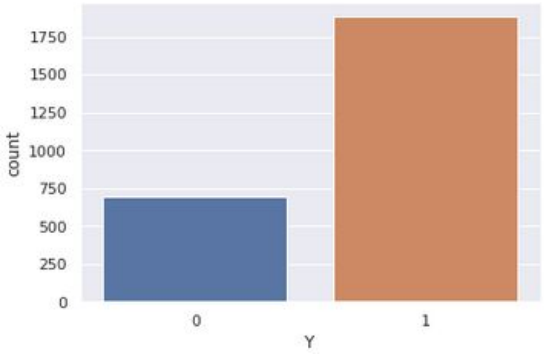
P4: threshold optimization

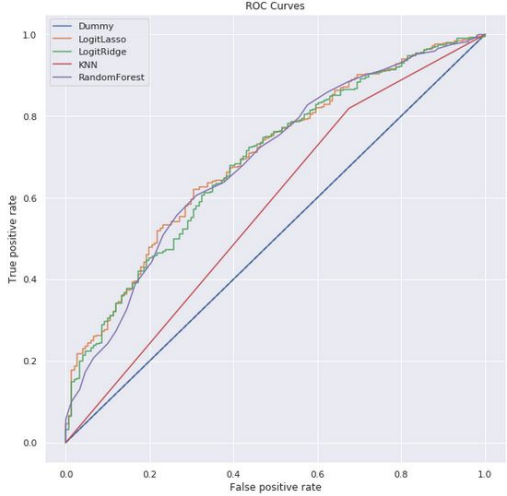
Results

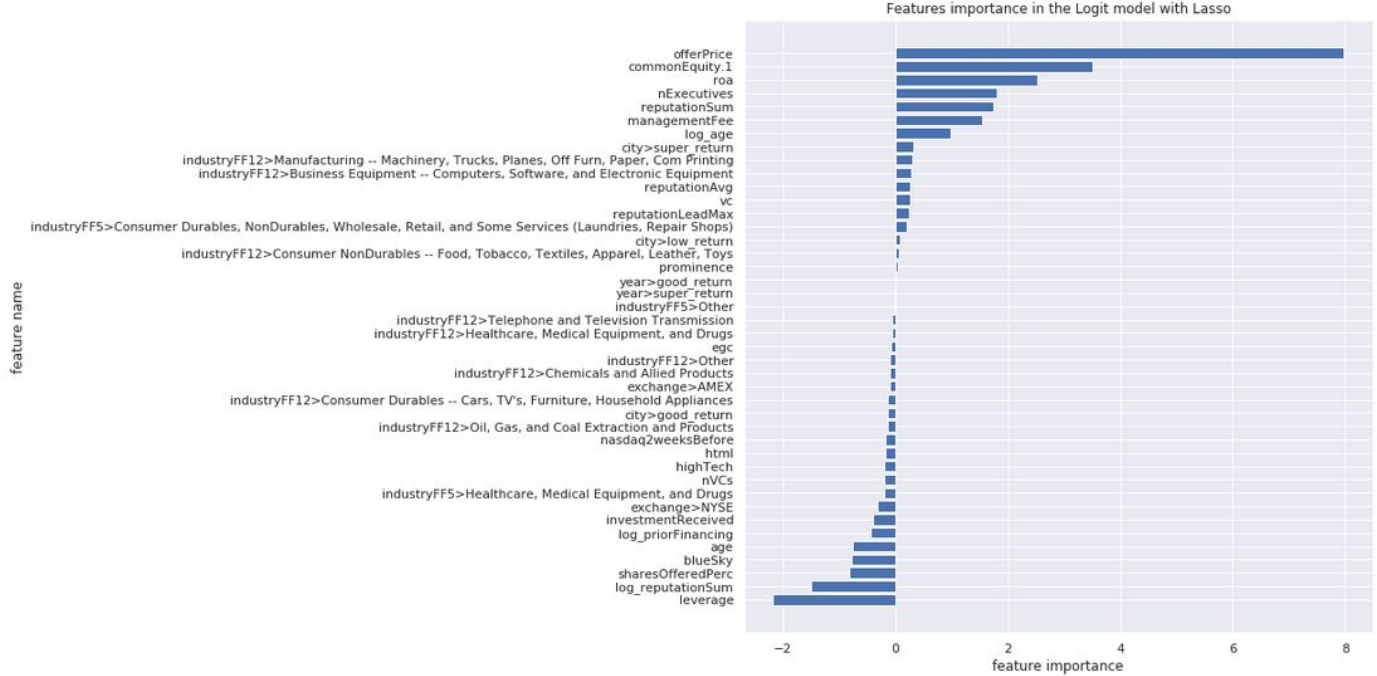
Problem	Model Retained	Class Imbalance	Performance Metric	Score	Optimal “TPR - FPR”
P1	Logit + LASSO	Mid	AUC	0.6994	0.285
P2	Logit 1 + Ridge	Mid	AUC	0.6054	-
P3	Logit + LASSO	Mid	AUC	0.6961	0.302
P4	RandomForest	Low	AUC	0.8170	0.522
P5	Logit + Ridge	Very high	AUC	0.6265	0.402
P6 (Regression)	RandomForest	-	MSE	101.6	-
P7	RandomForest	Low	P7 Custom score	1995.85	-
P8	RandomForest	High	P7 Custom score	765.20	-
P9	KNN	Very high	P9 Custom score	11.10	-

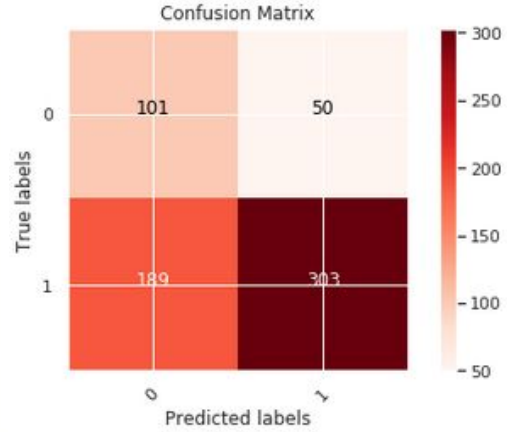
P1 :

Price change > 0 without rf



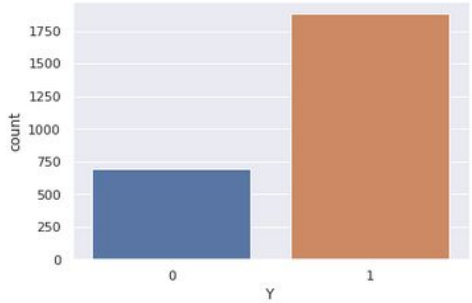


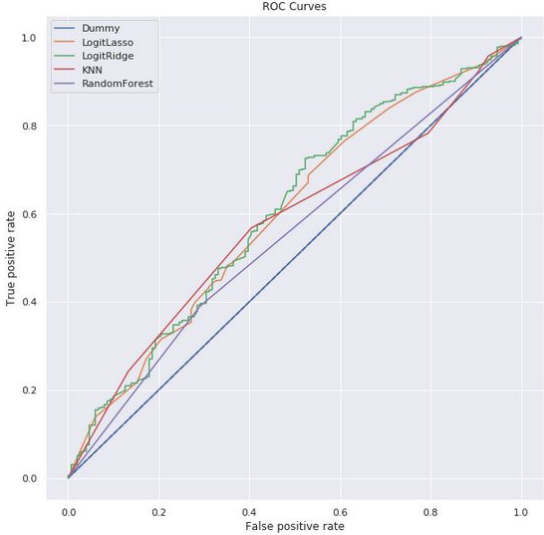


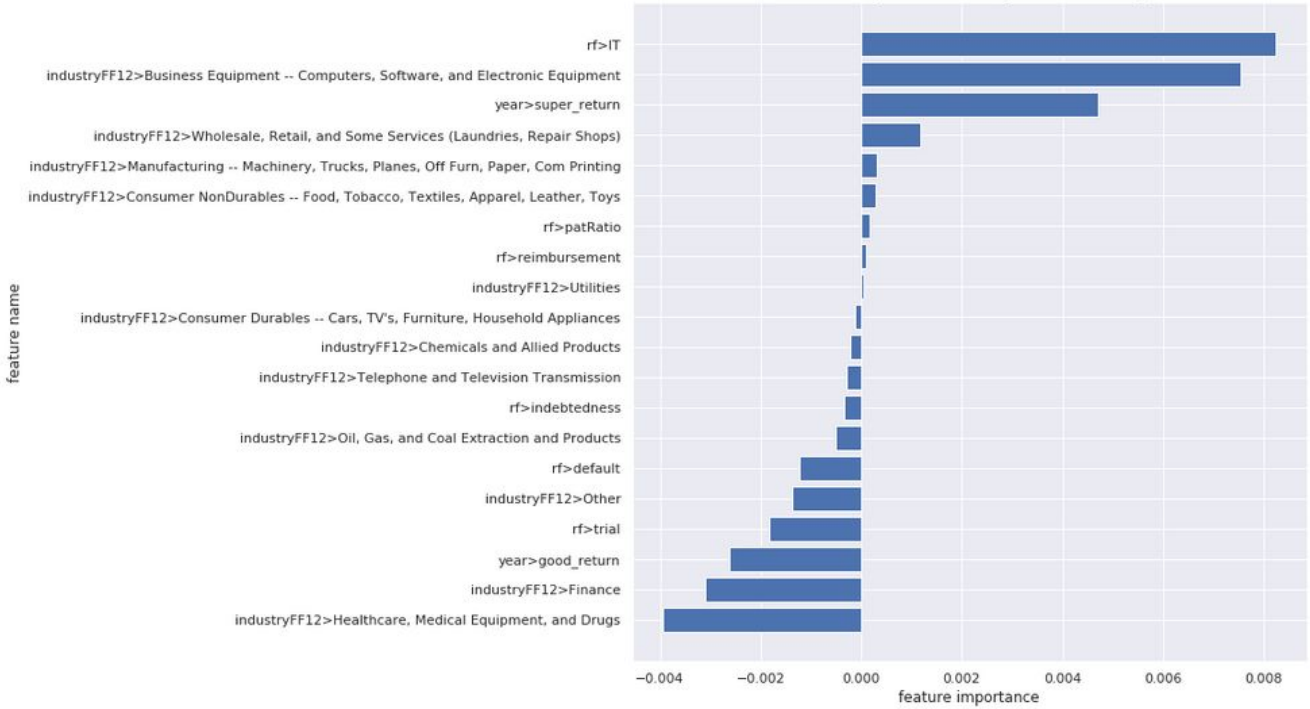


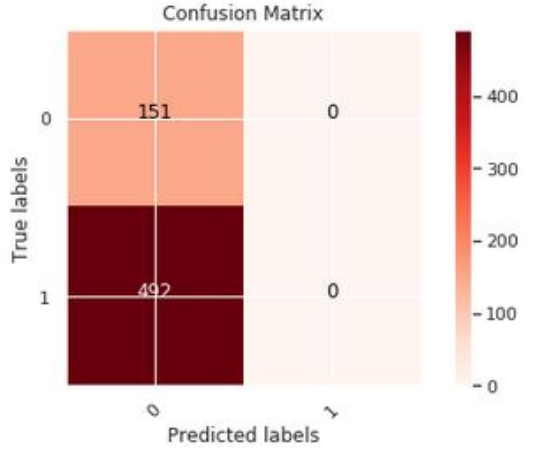
P2 :

Price change > 0 without rf, year and IndustryFF12



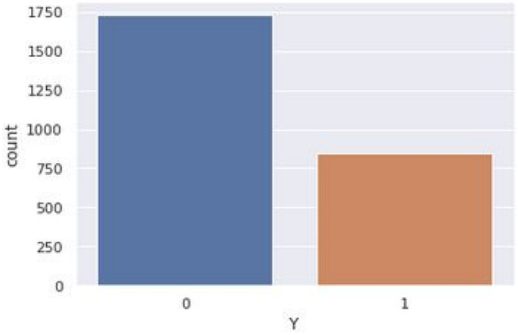


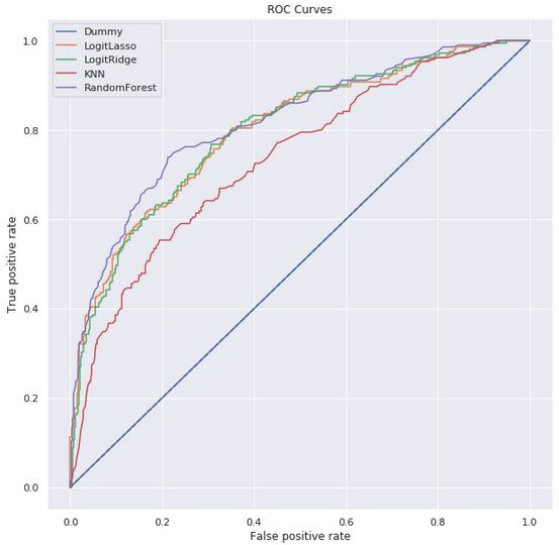


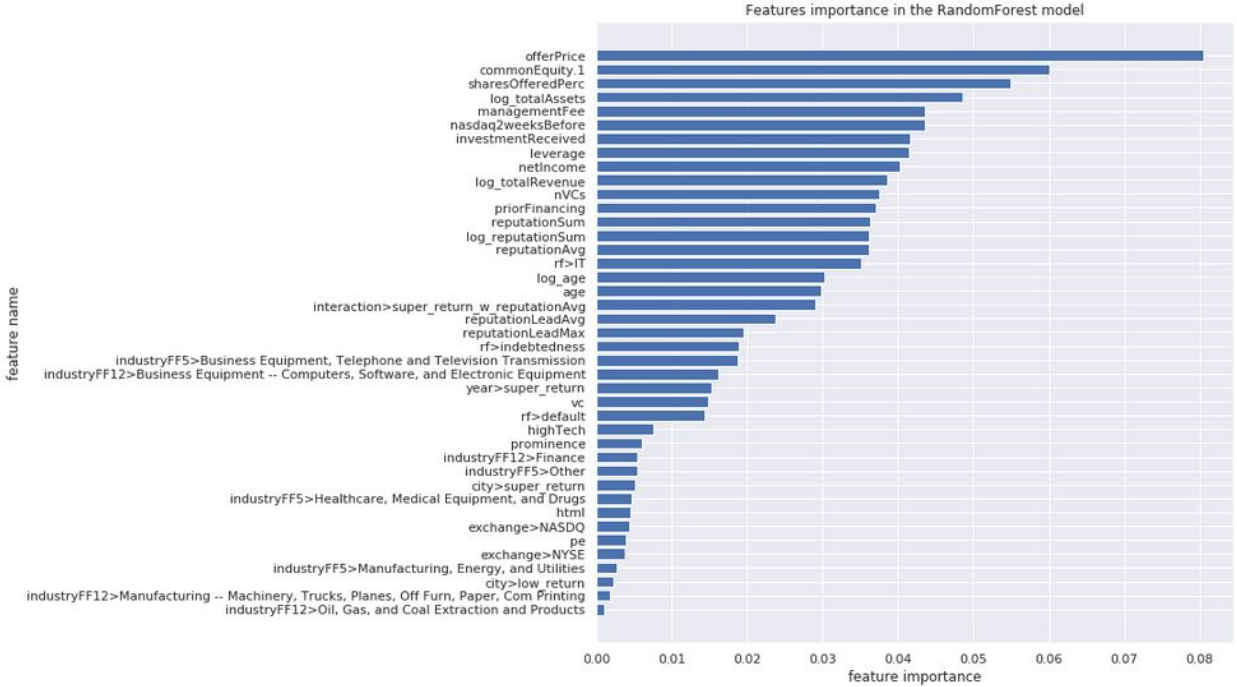


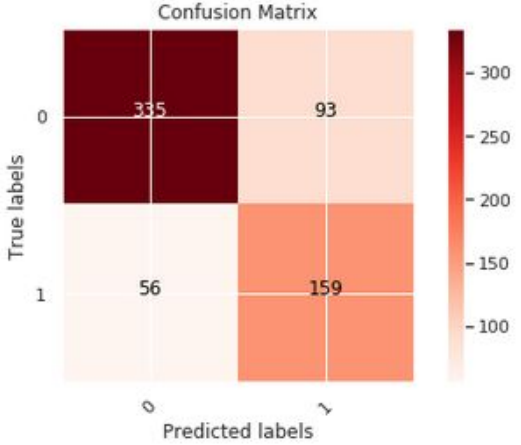
P4:

Price change > 20%



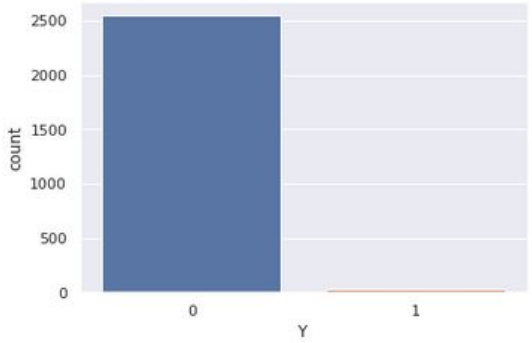


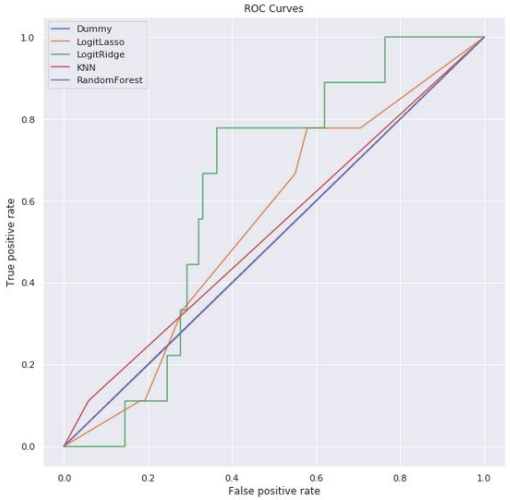


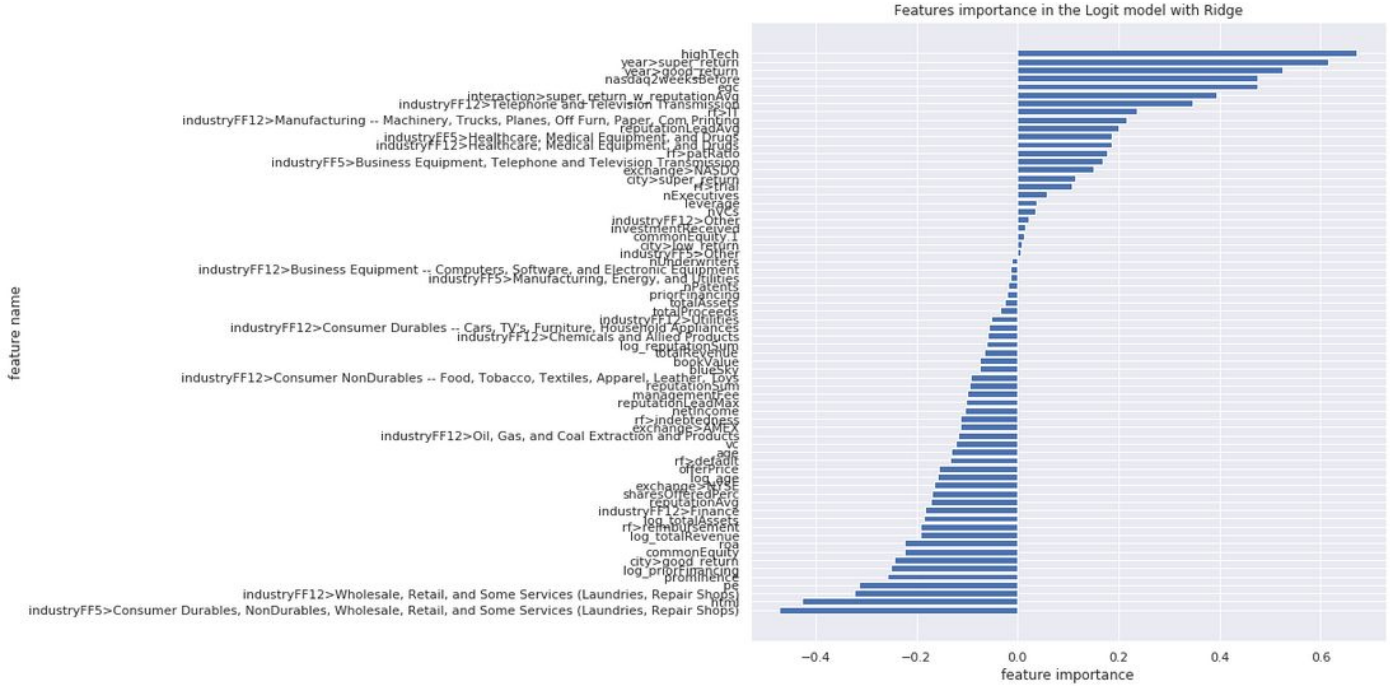


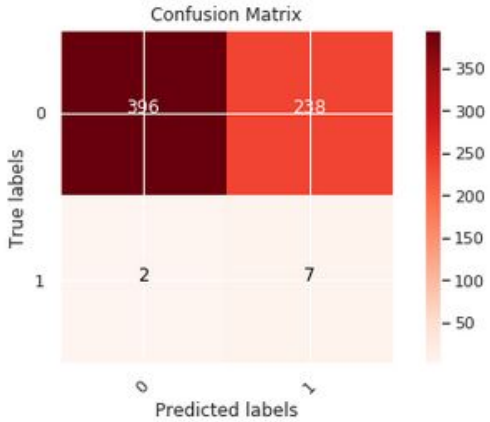
P5:

Price change < -20%

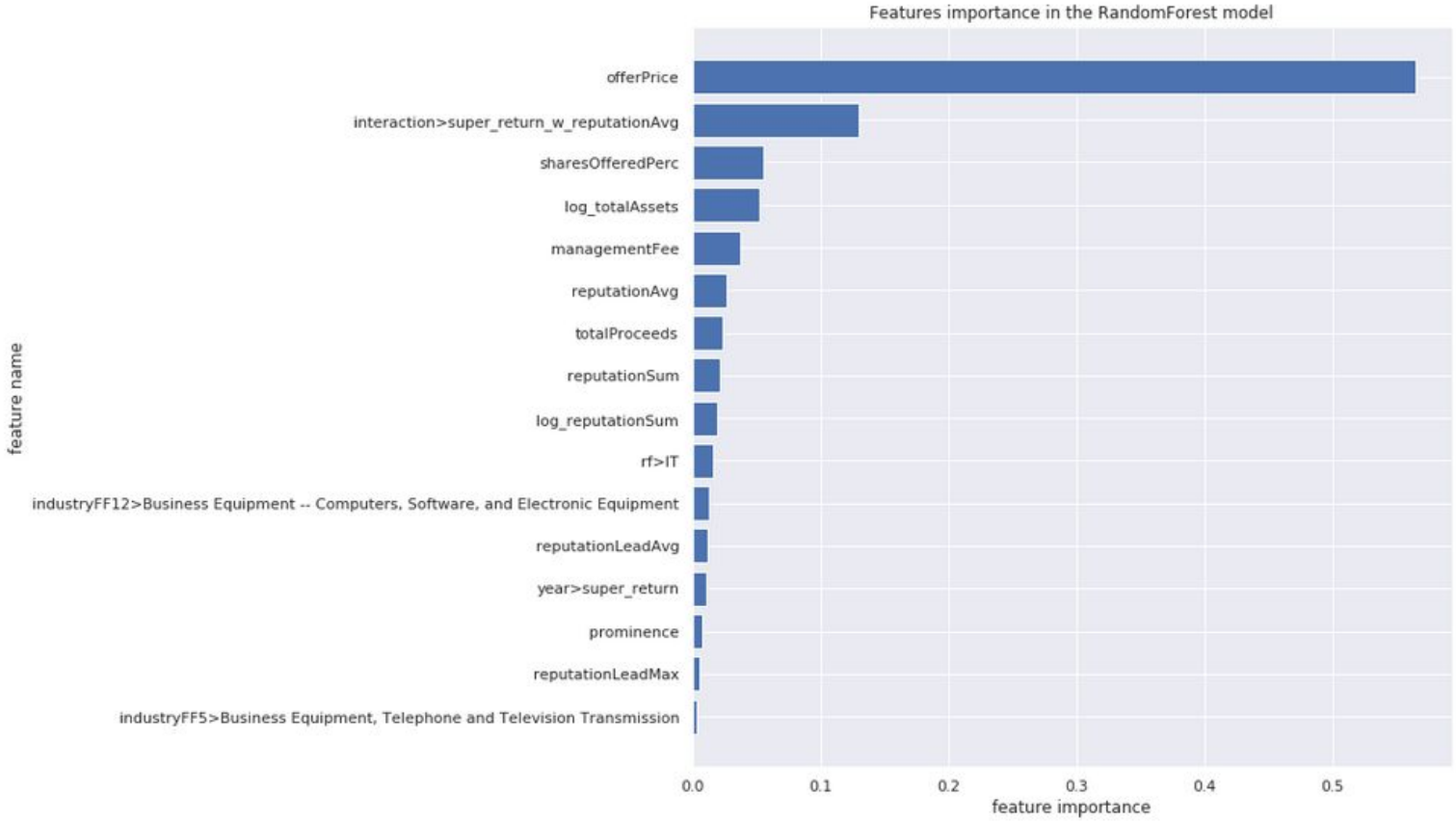






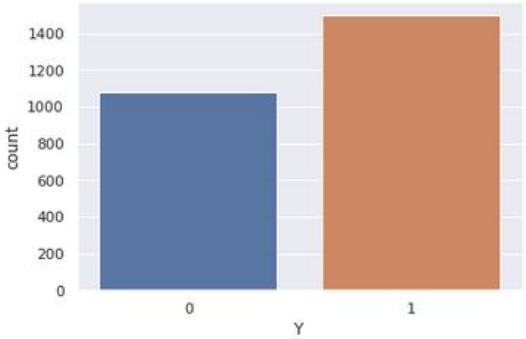


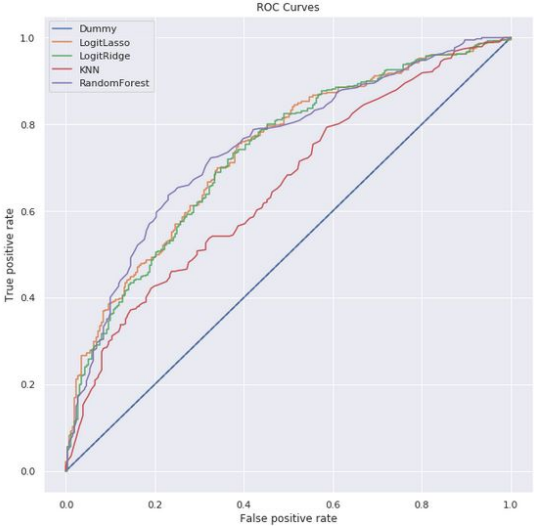
P6: Predict share price at the end of the first day

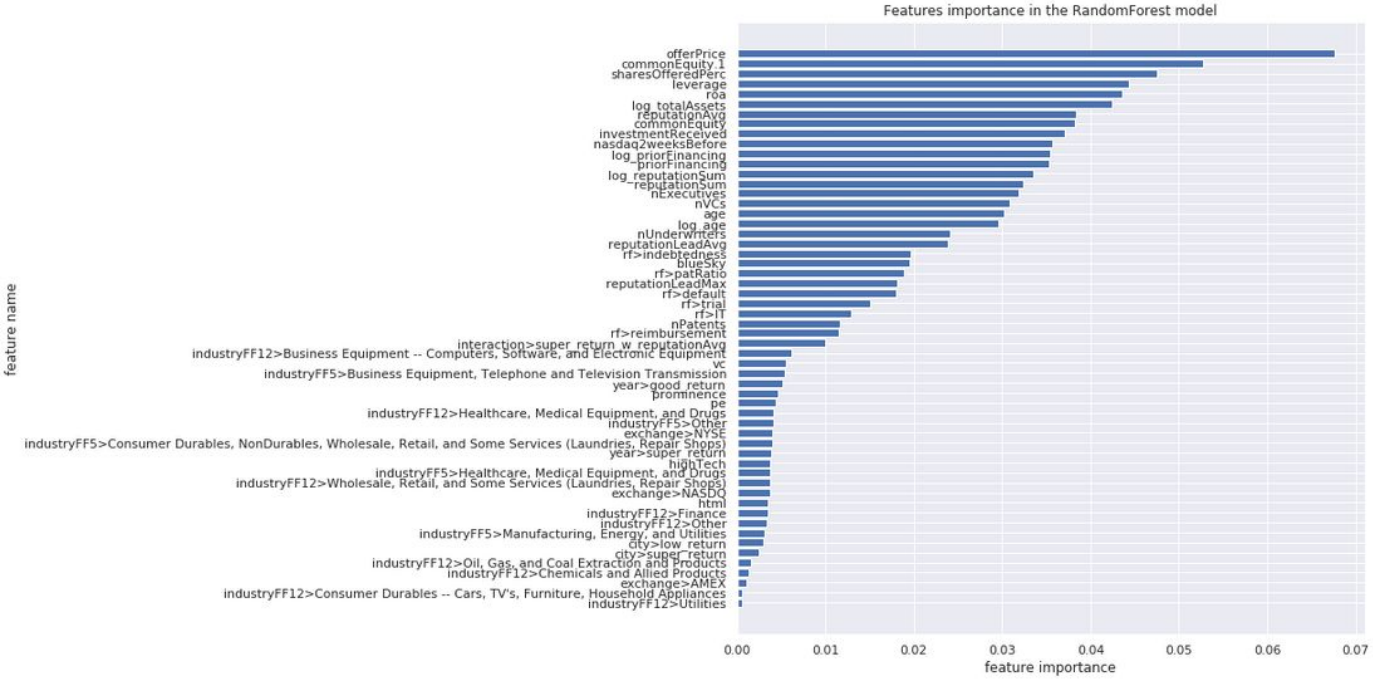


P7:

Price change > 5%

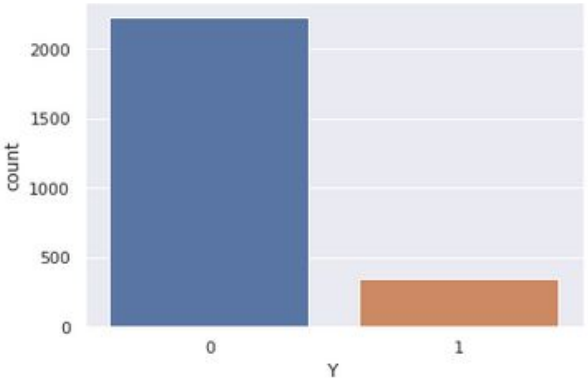


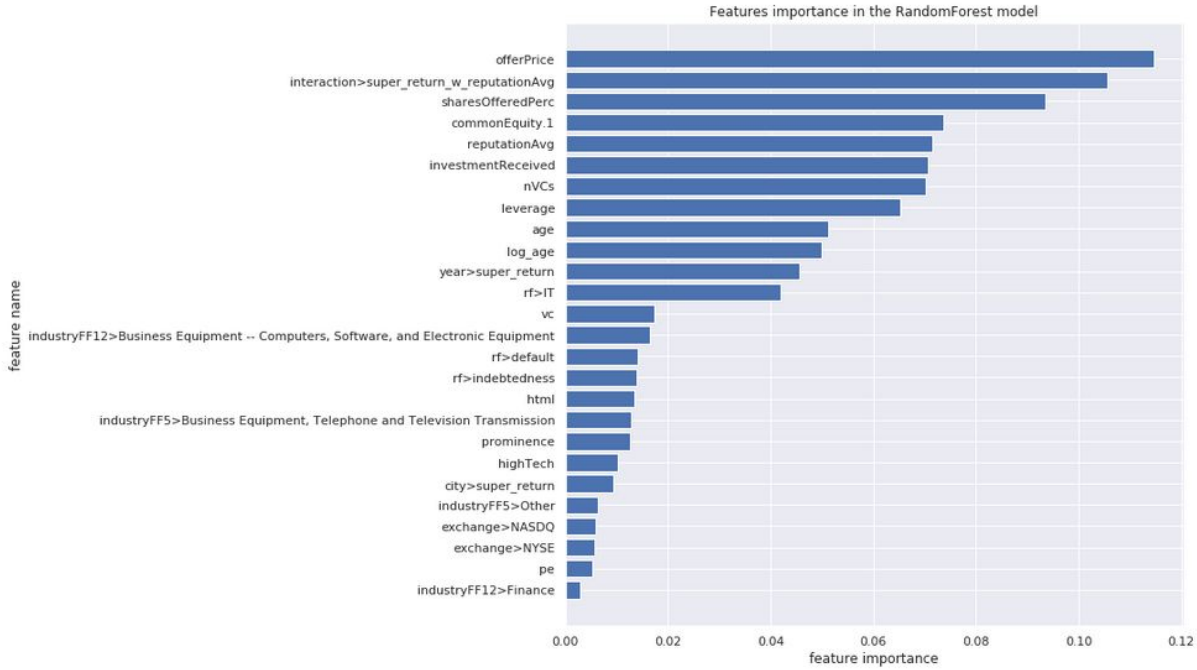


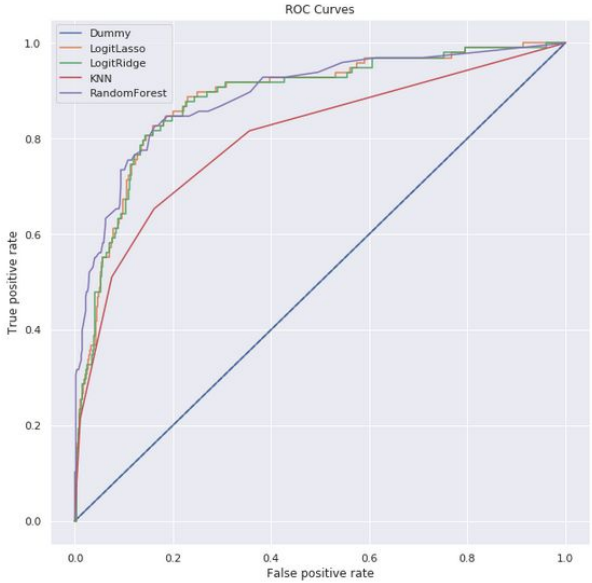


P8:

Price change > 50%







Thank you for your attention !

Questions