DATA SCIENCE FOR BUSINESS

course project: predicting IPO market valuation and price change

Dan Chai
Yaqiao Deng
18.01.2018

Content

- **Data Preprocessing**
  - Numeric feature, Categorical features, Text
- **Regression without Risk Factors**
  - Feature Selection
  - Predictive Models
- **Regression with Risk Factors**
  - Feature Selection
  - Predictive Models
- **Classification with Risk Factors**
  - Feature Selection
  - Predictive Models
- **Conclusion**
- **Additional Tasks**

Data Preprocessing

Numerical Features: Missing Values and Outliers Handling

- Keep the numeric features with more than 2500 non-NAN values
- Impute the missing values by the feature's median
- Winsorize the outliers (set the lowest value at 5% percentile and highest at 95% percentile)
- Log transformation for the positive features with long right tail

Categorical Features : Encoding

- Keep the categorical features with less than 10 unique values
- Encode 'yes/no' features into binary features (0 or 1)
- Encode categorical features taking more than two unique values into a series of 0 and 1

Text

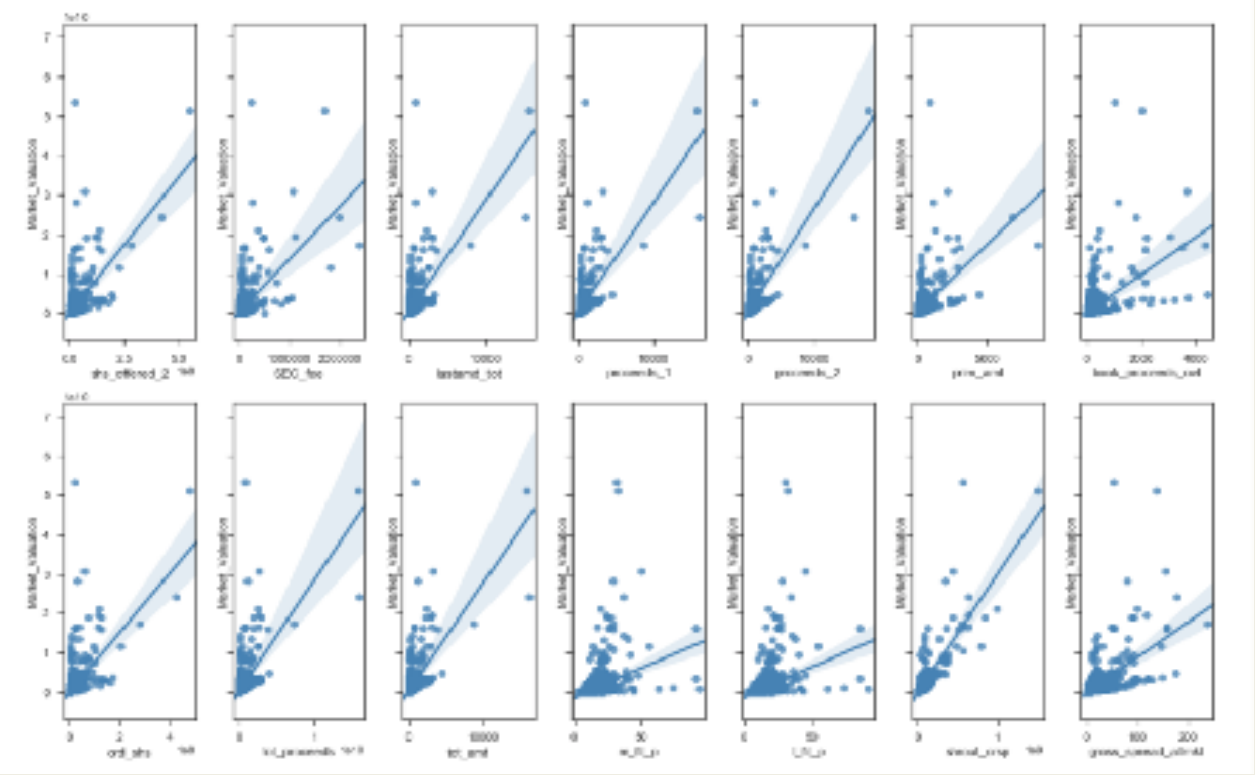
- Remove all extra white space and encoding symbols (\x94..)
- Convert words into lower case
- Get rid of stop words, numbers, special symbols and characters
- Tokenize and lammatize the words
- Keep the tokens appearing more than 100 counts in the corpus, resulting in 4270 unique words
- Create dictionary and bag of words
- TF-IDF transformation
- Use LSI and LDA to extract topics from the text

Regression without Risk Factors

Feature Selection

- select Kbest by f-regression
- select Kbest by mutual_info_regression
- select by feature importance with tree model

take the combine set of the selected features, which contains 14 features in total



All these selected features are originally numeric features

Regression without Risk Factors

Predictive models

- Linear Regression: Lasso, Ridge
- KNN Regression
- Decision Tree Regression
- Decision Tree with bagging method
- Decision Tree with random forest method
- Decision Tree with adaptive boosting method
- Support Vector Machine Regression

	MSE (e+18)	MAE (e+8)	MEDAE	R2
LINEAR	**2.69**	**3.18**	**8.15**	**0.73**
KNN	**3.40**	**3.50**	**7.61**	**0.52**
DECISION TREE	**3.61**	**3.50**	**5.46**	**0.72**
DT BAGGING	**2.52**	**2.84**	**4.92**	**0.63**
DT RANDOM FOREST	**3.12**	**2.83**	**4.98**	**0.51**
DT ADABOOST	**2.17**	**2.47**	**4.11**	**0.64**
SVR	**6.65**	**6.11**	**1.39**	**0.09**
BASELINE	**7.38**	**8.42**	**5.38**	**-0.0018**

Regression with Risk Factors

Predictive models

- Linear Regression: Lasso, Ridge
- KNN Regression
- Decision Tree Regression
- Decision Tree with bagging method
- Decision Tree with random forest method
- Decision Tree with adaptive boosting method
- ~~Support Vector Machine Regression~~

Use only LSI to extract topics from text

Feature Selection

- select Kbest by f-regression
- select Kbest by mutual_info_regression
- select by feature importance with tree model

take the combine set of the selected features, which contains 15 features in total

All these selected features are numeric features

Regression with Risk Factors

Model Comparisons

	MSE (e+18)	MAE (e+8)	MEDAE	R2
LINEAR	2.73	3.16	8.24	0.73
KNN	2.69	3.69	8.35	0.63
DECISION TREE	1.73	3.21	5.72	0.70
DT BAGGING	2.11	2.75	4.85	0.71
DT RANDOM FOREST	2.33	2.88	5.16	0.70
DT ADABOOST	1.46	2.71	4.98	0.79
BASELINE MODEL	7.38	8.42	5.38	-0.0018

Classification with Risk Factors

Feature Selection (LSI)

- select Kbest by f-regression
- select Kbest by mutual_info_regression
- select by feature importance with tree model

take the combine set of the selected features, which contains 19 features in total

Feature Selection (LDA)

- select Kbest by f-regression
- select Kbest by mutual_info_regression
- select by feature importance with tree model

take the combine set of the selected features, which also contains 19 features in total, but slightly different from LSI method

	Logistic (LSI)	Logistic (LDA)	KNN (LSI)	KNN (LDA)	SVM (LSI)	SVM (LDA)	XGBoost (LSI)	XGBoost (LDA)	Baseline model
accuracy	0.785	0.793	0.730	0.733	0.795	**0.800**	0.757	0.778	0.750

Conclusion

- We will use linear model with regularization in our final prediction without risk factors
- We will use decision tree with adaboost method in our final prediction with risk factors
- We will use SVM (using LDA to extract topics from text) in our final classification with risk factors
- It should be mentioned that this conclusion would be statistically more stable if we conduct several test sets, i.e. using k-fold cross validation across test sets.

Additional Tasks

- We only removed some extreme values on the right tail , maybe we could also find a way to remove the left tail (those are very small values), and aslo more methods to impute outliers
- We could try more methods for feature selection, so that categorical features might be selected
- We could do a nested cross-validation between the training and testing set
- We might try different numbers of topics (rather than 20) in the regression with risk factors, and try LDA model for the regression as well
- We could try to predict the market valuation according to the cosine similarity of the different risk factors. Or we could use PCA to extract some important components that are more relevant to our context.

THANK YOU

Any questions?

Question 1:

'proceeds_1', 'ord_shs', 'shs_offered_2', 'gross_spread_allmkt', 'm_fil_p', 'lastamd_tot', 'proceeds_2', 'tot_amt', 'book_proceeds_ovt', 'tot_proceeds', 'l_fil_p', 'shrout_crsp', 'SEC_fee', 'prim_amt'

Question 2: (LSI)

'print_exp', 'lastamd_tot', 'book_proceeds', 'shs_offered_2', 'proceeds_2', 'tot_proceeds', 'proceeds_1', 'amd_lp', 'tot_amt', 'gross_spread_allmkt', 'shrout_crsp', 'ord_shs', 'amd_mp', 'prim_amt', 'h_fil_p'

Question 3: (LSI)

'lsi3', 'quiet_period', 'lsi9', 'shs_offered_2', 'pctchg_lp', 'proceeds_2', 'lsi2', 'days_in_registr', 'num_amd', 'pctchg_nasdaq_1', 'gross_spread_allmkt', 'lsi16', 'pctchg_hp', 'filing_date', 'book_proceeds_ovt', 'shrout_crsp', 'date_amd', 'pctchg_mp'

Question 3: (LDA)

'lda16', 'quiet_period', 'lda6', 'shs_offered_2', 'pctchg_lp', 'proceeds_2', 'lda4', 'num_amd', 'pctchg_nasdaq_1', 'lda17', 'gross_spread_allmkt', 'SP1', 'pctchg_hp', 'filing_date', 'book_proceeds_ovt', 'date_amd', 'lda2', 'lda11', 'pctchg_mp']