A machine learning model for startup selection

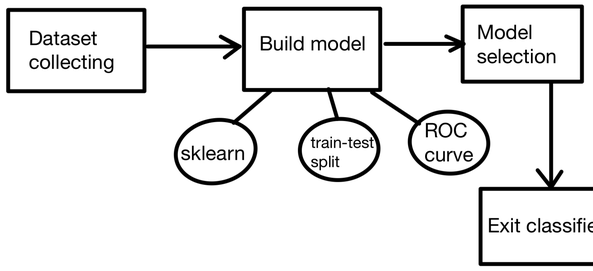
Khadjimurad Gamzatov
Optimization Class Project. MIPT

Introduction

According to an article [1], whose author cites Gatner and say that by 2025, 75% of venture capital funds will use AI to make investment decisions. There is also a Harvard study [2] that compared the ml built model with 200 venture angels. The ML model performed more profitably than novice or moderately experienced investors, but worse than very experienced ones. But still useful because the model finds insiders and the hybrid approach is the most profitable. Let's develop our ML model on publicly available data, look at different algorithms and the metric values they produce, and choose the best one.

Architecture of solution

Our system, as with most machine learning systems, starts with the data. We settled on manually exporting data from Crunchbase [3].To our features we apply a train-test-split and then test out various machine learning models and choose model with highest accuracy.



Data

This data comprises of 100K rows of people data and 100K rows of company data, both sorted by Crunchbase rank. We then joined these two datasets on the 'Primary Organization' column, which lead to about 60K rows of data. This dataset had to be filtered further for only founders, both successful and not, so we filtered the 'Job Title' column for those that contained only 'founder' or 'ceo'. We then dropped NaN values for the columns left over, of which there were many, which lead us to have 25K rows of data remaining. Once this process is completed, we selected various features that we wanted and recoded them to categorical or numerical. We also create an independent output variable, which is binary for whether the startup and founder raised at least a Series B.

The data contain the following fields 'Full Name', 'Primary Job Title', 'Bio', 'Gender', 'Number of News Articles', 'Number of Founded Organizations', 'Number of Portfolio Companies', 'Number of Investments', 'Number of Partner Investments', 'Number of Lead Investments', 'Number of Exits', 'Number of Events', 'Categories', 'Headquarters Location ', 'Operating Status', 'Founded Date', 'Closed Date', 'Company Type', 'Number of Founders', 'Success'

Also we Incorporated normalization to reduce overfitting and improve accuracy. Since HQ Location is a categorical var. with over 1000 levels, we changed these values to the frequencies of that particular location. Primary Job Title column had couple of different titles for each person (such as "CTO, co-founder, Product Manager"). Convert these to a binary variable that shows whether the person is the founder or not. And the RoBERTa model was used for texts [4].

Models and learning

To our features we apply a train-test-split (80/20) and then test out XGBoost, Random Forest, Logistic regression, SVM, kNN.

:

- t_p - we predicted the positive label and guessed
- f_p - we predicted a positive label, but were wrong in our prediction
- t_n - we predicted the negative label and guessed
- f_n - we predicted a negative label, but we were wrong

$$true\_positive\_rate = \frac{t_p}{t_p + f_n}$$

$$false\_positive\_rate = \frac{f_p}{f_p + t_n}$$

On tpr and fpr the ROC curve will be plotted.

$$Precision = \frac{t_p}{t_p + f_p}$$

$$Recall = \frac{t_p}{t_p + f_n}$$

On $Precision$ and $Recall$ we will build an Average Precision curve

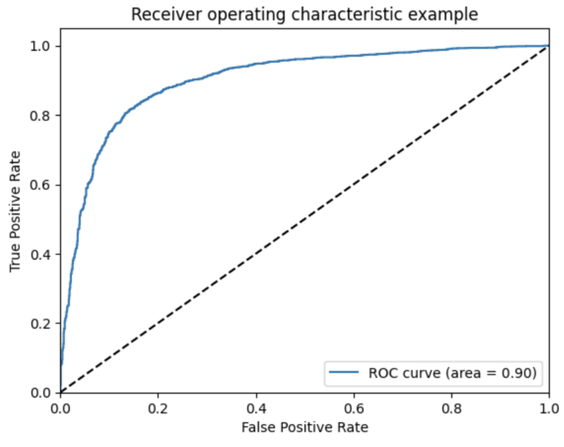
Accuracy table

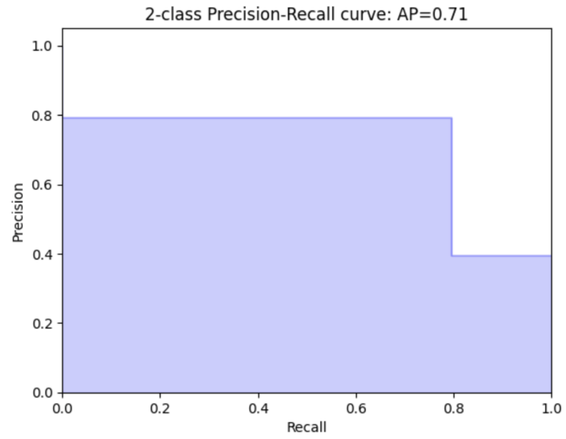
Below is a table with the accuracy metric calculated for all trained models, for training data and test data

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

	XGBoost	Random Forest	Logistic Regression	SVM	kNN
Train Accuracy	99.95	99.95	60.94	71.94	89.08
Test Accuracy	83.74	80.58	60.56	71.56	76.33

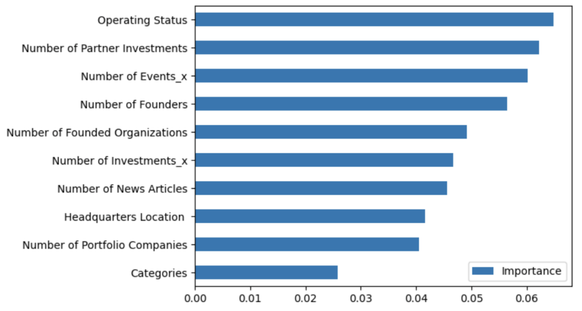
You can see that XGBoost did the best. Let's plot the ROC-curve and the AP-curve.





Results

Different models were trained and the best of them according to the accuracy metric was selected. We will also display a graph of the importance of the features for the selected model.



The code and models are in the GitHub repository [5].

References

[1] Kyle Wiggers. Gartner: 75% of vcs will use ai to make investment decisions by 2025. *VentureBeat*, 2021.

[2] Charlotta Siren Dietmar Grichnik Malin Malmstrom Torben Antretter, Ivo Blohm and Joakim Wincent. Do algorithms make better — and fairer — investments than angel investors? *Harward Business Review*, 2020.

[3] Crunchbase datasets: https://data.crunchbase.com/docs.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Github: https://github.com/murickg?tab=repositories.