Machine Learning Engineer Nanodegree
Capstone Project Proposal

Startup's success forecasting

Fernando Leandro dos Santos

May 16, 2017

1 Domain Background

Startup companies are recently emerging and evolving more than ever. But still, many of them fail, in fact, 9 out of 10 startups fail [1]. So, what are the characteristics of startups that succeed? There are many factors that might play a role in this question: founders and team experience, fundings, location, product, market and so on. Why some startups are acquired very early and others are not? Are there maybe some factors that attract bigger companies into some startup acquisition? This study aims to dig into these questions by investigating a dataset of startup companies and acquisitions provided by Crunchbase[2], a company that provides business information with data about companies and investments all over the world.

Previous researchers have tried to predict startups success from different types of data. Some have tried to predict startup acquisitions based on similar data (Crunchbase) together with TechCrunch[3] news about the companies[3]. Others have tried to predict startups survival time based on a venture screening questionnaire[2] as well as predicting success from business model documents[1].

2 Problem Statement

The startup world requires a lot of decision making from the people involved, specially when dealing with situations of funding or acquisitions that will impact in ownership changes of the company. In these situations, the decision maker should have on hand all information that can be provided and with the amount of data available today about companies, funding rounds and acquisitions, it could be very helpful to have statistical models helping and optimizing this decision. This study proposes a classification model that tries to predict if any operating company has more chances to be acquired (indicating success) or to be closed (indicating failure) in a near future.

3 Datasets and Inputs

The dataset used in this project is publicly provided by Crunchbase[4] for non-commercial use in accordance with the Creative Commons Attribution License requirements. The complete dataset contains information about organizations, people, offices, IPOs, investments, funds, funding rounds and acquisitions. For this project, most of this data will not be used. The data used here regards simply about organizations (companies only) and whether they were acquired, closed or were still operating by the time when the data were collected.

The initial data for this project was gathered from Peter Tripp's GitHub repository[5]. The dataset in this repository was already extracted from Crunchbase dump file into CSV files and dates from December 4, 2015. The dataset contains information of 66.368 companies from many countries, but only USA based companies will be considered for this study, which will decrease the amount of companies to 37.598 (this number might change after pre-processing and cleansing of the data).

[1] http://fortune.com/2014/09/25/why-startups-fail-according-to-their-founders/

[2] https://www.crunchbase.com/

[3] https://techcrunch.com/

[4] https://www.crunchbase.com/

[5] https://github.com/notpeter/crunchbase-data

The following attributes are directly available for each company: *name, url, company acting fields, total funding received, funding rounds, country, state, region, city, founding date, first funding date, last funding date and current status (closed, operating, acquired).* More attributes can be derived, if necessary, from other datasets, like investments, rounds or acquisitions.

4 Solution Statement

This study proposes a supervised learning approach to create a model that might be able to forecast a startup success as well as a data exploration approach to better understand the variables that most impact on the acquisition or failure of a startup company. Different supervised learning algorithms will be tested in order to find the most appropriate and the one with better performance according to the evaluation metrics defined below.

For exploring the data and better understanding of the features, some kind of PCA algorithm will be used. That might reveal the most expressing features for predicting a startup acquisition. After that, some kind of tree-based model will be applied to check in more detail these expressing features and which range of values are playing a more important role for the classification task.

Once the data is well explored and the best features are selected, Support Vector Machine and K-Nearest Neighbor algorithms are gonna be modelled. Since we'll have a relatively large amount of samples (more than 30.000), applying SVM will not be a problem. These algorithms were chosen due to their extra flexibility by testing different Kernels (for SVM) as well as different distance functions (for KNN). Also, those are from two different family of algorithms, kernel methods and instance based, so they are covering a good range of approaches to deal with our classification problem.

5 Benchmark Model and Evaluation Metrics

The model proposed by Xiang et al.[3] will be used as a comparison base. They created a model using Crunchbase data as well, but from a different period of time. They measured their model performance using a True Positive rate, False Positive rate and the area under the ROC curve (AUC). It is worth noting that they included in their model data from TechCrunch news articles about the companies, which they concluded to have improved the overall performance of the model. In this project, the same metrics will be used to evaluate the model.

True Positive and False Positive rates are fairly simple metrics, directly derived from a classification confusion matrix of a binary classifier:

- **True Positive (TP)**: Rate of instances that were classified with a positive label and are positive
- **False Positive (FP)**: Rate of instances that were classified with a positive label and are **not** positive
- **True Negative (TN)**: Rate of instances that were classified with a negative label and are negative
- **False Negative (FN)**: Rate of instances that were classified with a negative label and are **not** negative

From these concepts, we can explain the intuition behind the AUC metric. The ROC curve is basically computed by combining TP rate with TN rate in such a way that by plotting the TPR (in the Y axis) and FPR (in the X axis), we'll have a curve as stated in Figure 1. The blue area corresponds to the area under the ROC curve while the dashed line represents the ROC curve of a random predictor, which serves as a baseline to see whether the model is useful.

6 Project Design

The implementation of this project will accomplish, in order, the following procedures of a workflow:

6.1 Data Preparation

In this part of the project, the dataset will be filtered, cleaned and mounted in order to simplify future analysis. This is where the the proper dataset will be prepared. The information about each company will be retrieved from all the files we have in the dataset and merged into one single large table. Companies with important missing attributes as well as not USA based companies will be removed and variables will be normalized and expanded to the simplest possible structure.

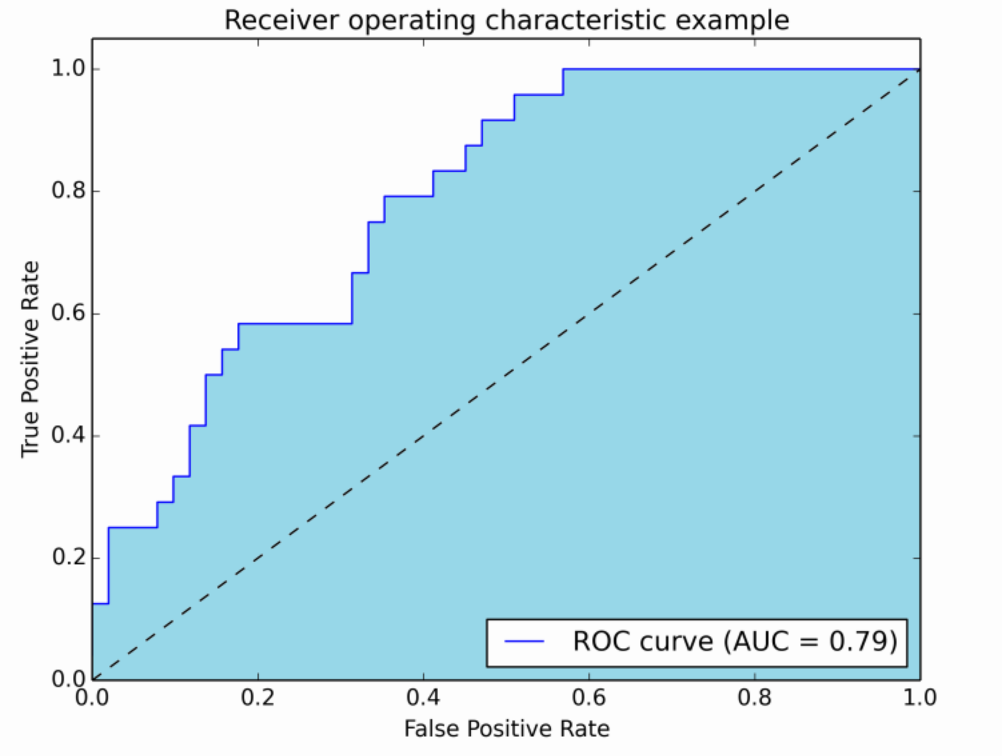

Figure 1: Area under the ROC curve in blue

6.2 Data Exploration

Once all the data is merged together in a single table and all possible attributes are in place, we start the data exploration. In this phase, attributes will be tested in order to discover their significance against forecasting the target variable. This phase will make use of graphical visualizations to explore and better understand the distribution of the variables. Also, algorithms like Principal Component Analysis (PCA) will be applied to the dataset in order to find the best attributes to be used in the Modelling phase.

6.3 Modelling

In this phase, a proper dataset with expressing input variables and a target variable will be already available. A few supervised learning algorithms will be applied and different parametrizations will be tested by a GridSearch technique. Depending on the amount of data and the algorithms used, K-fold cross validation technique might also be used.

6.4 Evaluation and Conclusion

Once the best model is found, it will be tested against a totally unseen part of the dataset. These final results will be analyzed and compared against a benchmark model. After that, a proper discussion will be made considering the effectiveness of the final model in a real world situation.

References

[1] Markus Böhm et al. “The Business Model DNA: Towards an Approach for Predicting Business Model Success”. In: (2017).

[2] WilliamB Gartner, JenniferA Starr, and Subodh Bhat. “Predicting new venture survival: an analysis of “anatomy of a start-up.” cases from Inc. Magazine”. In: *Journal of Business Venturing* 14.2 (1999), pp. 215–232.

[3] Guang Xiang et al. “A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch.” In: *ICWSM*. 2012.