Startup success

Our goal is to determine which factors have effect on startup success. Data includes 472 startups. Every startup has 115 features and one label column, which is show was startup succeeded or failed.

Data manipulation

Data have many missing values, that why we should use some technics to manipulate data.
I try to don't lose any startup, that's why, for continues variables I fill missing values with mean or median. For categorical variables missing values was filled with most appeared value, and after that I make them dummies variables. After appending all variables in one data frame, our clean data frame has 214 columns.

Data preparation

Because our data features have different scales, I normalize data to train models more efficient. Because we have lot of features, I decide to use some dimension reduction technique, especially PCA. I take first 25 principal components, which are explain 50% variance of data. I split data into training set (80%) and test set (20%).

Logistic model

I build Logistic regression model on PCA components and get 90% accuracy for test set. This is the model summary.

	precision	recall	f1-score	support
0	0.87	0.84	0.85	31
1	0.92	0.94	0.93	64
accuracy			0.91	95
macro avg	0.89	0.89	0.89	95
weighted avg	0.90	0.91	0.90	95

Just to remember formulas.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

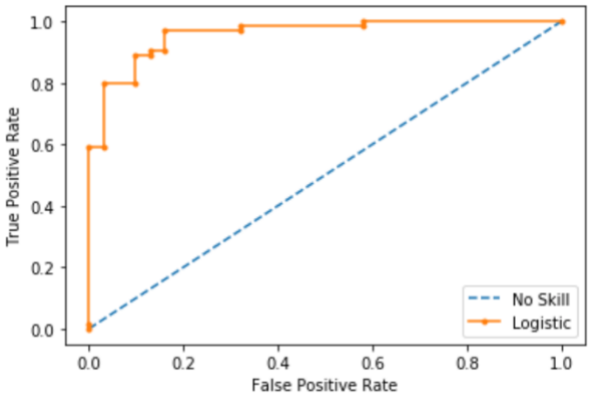
TP = True positive

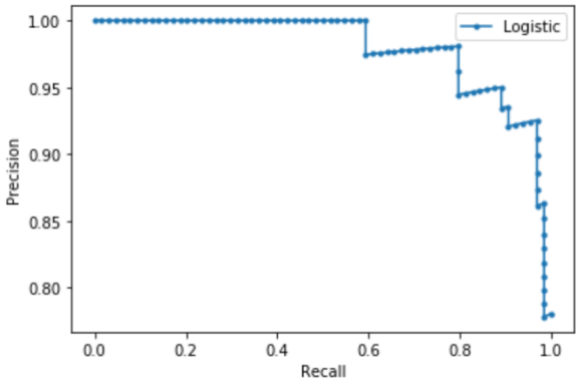
TN = True negative

FP = False positive

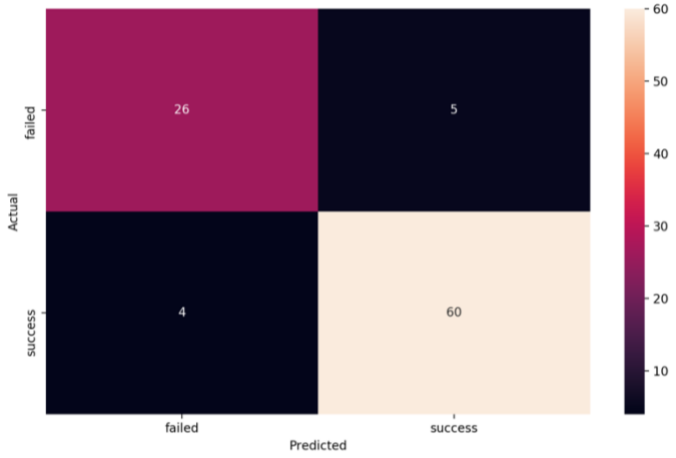
FN = False negative

Here is the ROC-AUC and PRECISION-RECALL curves.
Area under the curve = 95

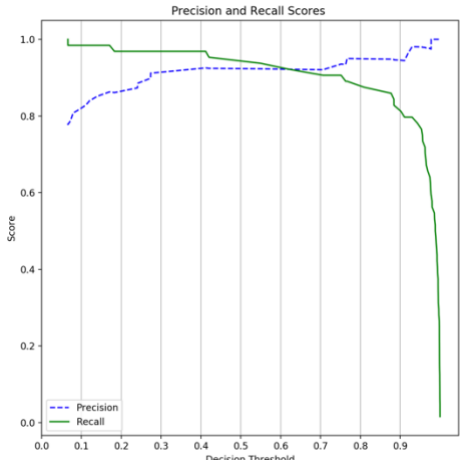




This is the confusion matrix



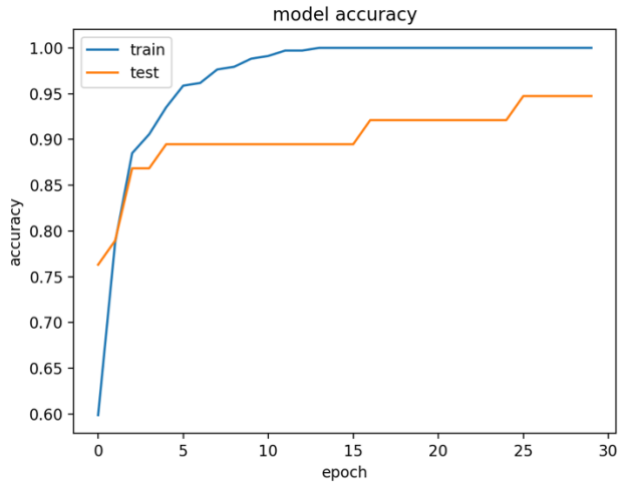
I decide to change threshold from 0.5 to another value to get more accuracy. I made decision using **PRECISION and RECALL** curves. I choose threshold = 0.6. Using this threshold get model accuracy=92.6%

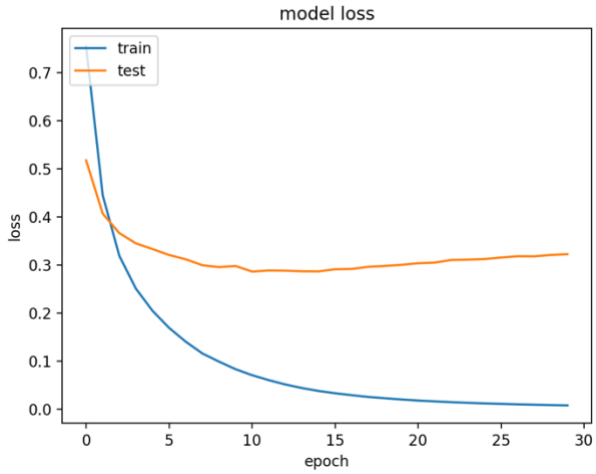


I also try 20-fold cross validation and get accuracy=0.87.

Multilayer Perceptron

I also try to construct some Neural Network model. Because data is small, I construct some small model. For first layer it has 20 neurons with activation function RELU. And on last layer I used SoftMax layer. I get 87% accuracy. Here is the learning process.





Reasons behind success

To explain what reasons are behind success, I take first principal component and calculate correlation with data features. If correlation is greater than 0.4, I consider that this variable is have an effect for success. These are these features.

	PCA1
PCA1	1.000000
Big Data Business_yes	0.554186
Technical proficiencies to analyse and interpret unstructured data_yes	0.502946
Proprietary or patent position (competitive position)_yes	0.492858
Experience in Fortune 500 organizations_1	0.467022
Experience in Fortune 1000 organizations_1	0.466598
Catering to product/service across verticals_yes	0.457776
Number of of Research publications_many	0.448881
Local or global player_global	0.447301
Experience in selling and building products_high	0.446136
Company awards_yes	0.445715
Dificulty of Obtaining Work force_medium	0.444813
B2C or B2B venture?_b2b	0.443418
grown	0.438045
Client Reputation_high	0.437957
Top management similarity_medium	0.424956
Predictive Analytics business_yes	0.418242

Conclusion

So, if we construct models like mentioned above, we can predict failure or success with about 90% accuracy and this is good, if we want to predict company success probability.
If we want to startup have success, according this report we should start an analytic business, like a Big Data business. To have success, businesses should grow their staff, should get experiences in selling and building products, should have good management system, work more with another businesses, should have good relationship with clients.