Start-up Pulse

Plan of Action

September 18, 2017

- Data-Preparation
- Exploratory DataAnalysis (EDA) - very Important, since we have a ton of features
- Feature Selection
- Model Selection

1 Suggested Flow for PULSE

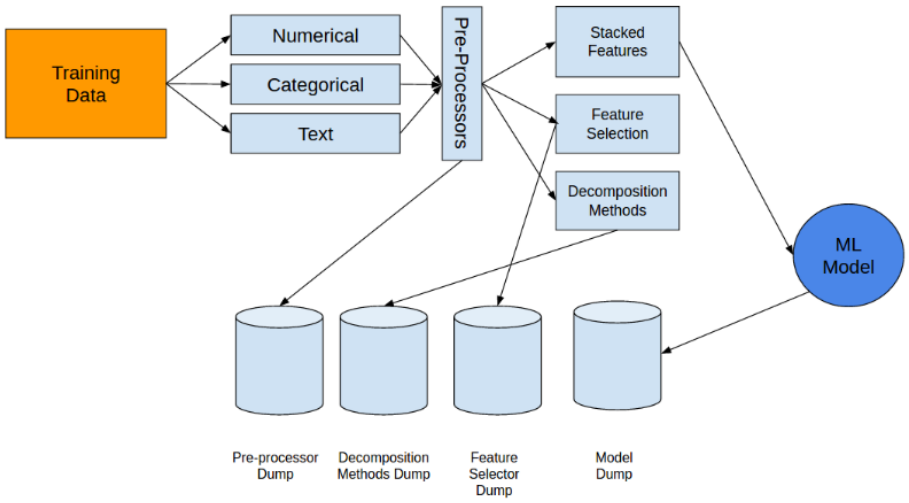

Figure 1: Suggested flow for start-up pulse

2 Data Preperation

Before applying the machine learning models, the data must be converted to suitable form.

- numpy
- pandas - Data Munging

2.1 label encoding

It is necessary to convert the categorical data into numerical entities so that it can be fed into the Model later

- using Dummy variables
- Please NOTE - For a given attribute variable, none of the dummy variables constructed can be redundant. That is, one dummy variable can not be a constant multiple or a simple linear relation of another.

2.2 Avoiding multicollinearity

Since the start-up data has huge number of features, it is very important to avoid multicollinearity in order to achieve better performance.

- I think that an elastic-net/ridge regression approach should allow you to deal with collinear predictors. I've included ridge regression in the Suggested Regression Algorithms section.

2.3 Few Suggestions

- Replace the numeric missing values (NaN's) with the mean of their respective columns
- transforming the skewed numeric features by taking log(feature + 1) - this will make the features more normal

3 Exploratory Data Analysis

Since the startup data has so many features, it is really important to understand even the subtle details and relationships between features. The following are the Suggested tools for EDA

- Plotly/Cufflinks - Best for Interactive plots
- Bokeh - Power to create custom visualizations
- Seaborn - Importantly PairPlots

4 Feature Selection

4.1 RFE - Recursive Feature Elimination

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.

NOTE - This helps us find the discriminating features in the startup-data

5 Suggested Regression Algorithms

Which parameters to optimize? How to choose parameters closest to the best ones? These are a couple of questions that can be answered only we have a clear look at the startup data. One cannot get answers to these questions without experience with different models + parameters on a large number of datasets.

I've tried to break down the hyperparameters, model wise:

Table 1: Suggested Algorithms for **start-up pulse**

Algorithm	Parameters to optimise	Good Range of Values
RandomForest Regressor	1.N_estimators 2.Max depth 3.Min sample split 4.Max Features	120, 300, 500, 1800 5, 8, 15, 30, None 1, 2, 5, 10, 15, 100 Log2, sqrt, None
Ridge Regression	1.Alpha 2.Fit intercept 3.Normalize	0.01, 0.1, 1, 10, 100 True/False True/False
Lasso	1.Alpha 2.Normalize	0.1, 1, 10 True/False
AdaBoost Regressor	1.N-estimators 2.LearningRate 3.Loss	Subject to change

6 Resources

List of Resources used from preparing the Investors questionnaire to Model selection

- For investors questionnaire
  - Risk factors that scares investors - http://blog.gust.com/7-startup-high-risk-factors-that-scare-investors/
  - Important Factors Venture Capitalists Consider Before Investing https://www.entrepreneur.com/article/293159
  - 7 Factors For Deciding To Invest In A Startup https://www.forbes.com/sites/mariannehudson/2014/09/18/7-factors-for-deciding-to-invest-in-a-startup-or-not/#7ead21fc344b
- List of domains,industries

  http://www.startupranking.com/startup/register
- Customer retention metric

  http://www.selligent.com/content/customer-retention-measurement
- Top 10 Revenue models

  https://fi.co/insight/the-10-most-popular-startup-revenue-models

- Selecting the correct metric

  http://www.businessinsider.com/how-do-you-select-a-revenue-model-for-your-startup-2014-8?IR=T

- Customer Retention metrics:

  http://www.iamwire.com/2016/07/5-metrics-to-calculate-customer-retention/139872

- De-risking a startup (resource on risks)

  https://codingvc.com/how-to-de-risk-a-startup

- Data Preparation

  https://machinelearningmastery.com/prepare-data-machine-learning-python-scikit-learn/

- numpy Doccumentation

  https://docs.scipy.org/doc/numpy-1.13.0/reference/

- pandas Doccumentation

  https://pandas.pydata.org/pandas-docs/stable/

- Recursive Feature Elimination

  http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

- Random Forests

  https://en.wikipedia.org/wiki/Random_forest

- AdaBoost Regressor

  Y. Freund, R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting", 1995.