PROJECT - Identify winners from fundamental data

PROBLEM

Description

Using historical financial statement information for the S&P 500 companies, can we identify winners from losers.

Scope

Scope of this is to leverage an academic study from the University of Chicago for value investing using accounting fundamentals.

Audience

Anyone speculating based on fundamental data.

DATA

Source

https://www.kaggle.com/dgawlik/nyse

Content

Data spans 2010 to 2016 and contains historical financial information for the S&P 500 obtained from SEC filings as well as prices from the New York Stock Exchange (NYSE).

Observation

Data doesn't reflect corporate actions correctly which can potentially skew the findings. More effort may be needed to clean the data or select some stocks.

HYPOTHESIS

Success depends on acceptable prediction capability of identifying winners based on fundamental information.

PROJECT - Wine Quality

PROBLEM

Description

Using the physicochemical tests identify the quality of the wine.

Scope

Scope of the idea is to only use the physicochemical tests (ie., there is no data about grape types, brand, selling price, etc.) to identify the quality of the wine.

Audience

Wine enthusiasts who are open for alternate ways to identify wine quality.

DATA

Source

http://archive.ics.uci.edu/ml/datasets/Wine+Quality

Content

Data consists of around 6000 samples of red and white wine data containing 12 physicochemical tests (such as acidity, citric acid, residual sugar etc.,)

Observation

Feature selection is important and considerable effort is needed to understand the data.

HYPOTHESIS

Success depends on the feature selection and the usefulness of the features in identifying the quality of the wine.

PROJECT - Caravan Insurance Data

PROBLEM

Description

Using the customer data for insurance company, can we predict whether the customer buys an insurance or not. In addition, can we derive profile information on a typical insurance buyer.

Scope

Scope of the idea is to better understand the importance of feature selection when the training data is unbalanced, meaning only a small % of customers actually buy the insurance.

Audience

Useful for the company marketing to target the right customers.

DATA

Source

http://kdd.ics.uci.edu/databases/tic/tic.html

Content

Data was part of computational intelligence and learning 2000 challenge. Data consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes.

Observation

Feature selection is important, hence considerable effort is needed to understand the data.

HYPOTHESIS

Success depends on the feature identification and how reasonable the features are in predicting the customer who is likely to buy the insurance.