GREYATOM

Hackathon Projects

By

GREYATOM

Table of Content

How will this Hackathon work?

Day 1 - Saturday, 25th November 2017 - 9.30 AM - 6 PM

- Hackathon use case intro and Hackathon kick-off (Abhi)
- Based on the problem statements selected, conduct a couple of hours workshop to guide people through any new concept not covered as part of the overall training - GreyAtom speaker
- Give Participants a high level task breakup and estimated timelines to achieve each milestone
- Mentor hours - 30 mins with each of the 4 teams
- Share good sample capstone report templates and videos on how capstones should be presented
- End of Day 1 - Milestone check 1

Day 2 - Sunday 26th November 2017 9.30 AM - 6 PM

- People work through their solutions
- Mentor Hours - 30 mins with each of the 4 teams
- Mid Day - Milestone check 2
- Final Presentations & Evaluation

Proposed Problem Statements

Build a model to approve/ disapprove a loan for a prospective applicant based on his/her profiles

Problem Statement:

When a bank receives a loan application, based on the applicant's profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. Two types of risks are associated with the bank's decision –

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank

A predictive model developed on this data is expected to provide a bank manager guidance for making a decision whether to approve a loan to a prospective applicant based on his/her profiles.

Data Set Information:

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. Here is a link to the German Credit data. Some key attributes of the data set are

1. Age (numeric)
2. Sex (text: male, female)
3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)
4. Housing (text: own, rent, or free)
5. Saving accounts (text - little, moderate, quite rich, rich)

6. Checking account (numeric, in DM - Deutsch Mark)
7. Credit amount (numeric, in DM)
8. Duration (numeric, in month)
9. Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

Data Source

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

GREYATOM

Predict if client will subscribe to direct marketing campaign for a banking institution

Problem Statement:

The data is related to direct marketing campaigns of a Portuguese banking institution. Predict if client will subscribe.

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs). The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

Goal :- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Sources

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

https://data.world/uci/bank-marketing

GREYATOM

Build a model to compute probability of default for Taiwanese Credit Card Clients

Problem Statement

Build a model to compute probability of default for Taiwanese Credit Card Clients

Abstract: This research aimed at the case of customers default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.

Data Set Characteristics:	Multivariate	**Number of Instances:**	30000	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	24	**Date Donated**	2016-01-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	135593

Data Set

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 25 variables:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years

- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Predict whether a particular company will default within 5 years, given its financial statement data (numeric only)

Problem statement:

The Financial statement of a firm consists of critical information such as EBITDA of a company, which can be used to track and predict the future of a company. Although important information is also present in form of text in Notes to Financial Statement section but, here we are simply using the numeric information for the prediction purpose.

Data: The dataset is about bankruptcy prediction of Polish companies. The data was collected from Emerging Markets Information Service, which is a database containing information on emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

- Based on the collected data five classification cases were distinguished, that depends on the forecasting period:
  - 1stYear : the data contains financial rates from 1st year of the forecasting period and corresponding class label that indicates bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 represents bankrupted companies, 6756 firms that did not bankrupt in the forecasting period.
  - 2ndYear : the data contains financial rates from 2nd year of the forecasting period and corresponding class label that indicates bankruptcy status after 4 years. The data contains 10173 instances (financial statements), 400 represents bankrupted companies, 9773 firms that did not bankrupt in the forecasting period.
  - 3rdYear : the data contains financial rates from 3rd year of the forecasting period and corresponding class label that indicates bankruptcy status after 3 years. The data contains 10503 instances (financial statements), 495 represents bankrupted companies, 10008 firms that did not bankrupt in the forecasting period.
  - 4thYear : the data contains financial rates from 4th year of the forecasting period and corresponding class label that indicates bankruptcy status after 2 years. The data contains 9792 instances (financial statements), 515 represents bankrupted companies, 9277 firms that did not bankrupt in the forecasting period.
  - 5thYear : the data contains financial rates from 5th year of the forecasting period and corresponding class label that indicates bankruptcy status after 1 year. The data contains 5910 instances (financial statements), 410 represents bankrupted companies, 5500 firms that did not bankrupt in the forecasting period.Train/Test data can be taken as a mix of all 5 data instances.
- Each row represents a company
- There are missing values which will have to be dealt with during Data Preparation stage
- For classifying into train/test a mix of data can be taken from the 5 categories.
- Some key attributes:

1. X1 net profit / total assets
2. X2 total liabilities / total assets
3. X3 working capital / total assets
4. X4 current assets / short-term liabilities
5. X48 EBITDA (profit on operating activities - depreciation) / total assets
6. X49 EBITDA (profit on operating activities - depreciation) / sales

Data Source: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

Hackathon Teams

Team 1	Karan Mehta
	Neelkanth Mehta
	Rishabh Mishra
	Rohan Gupta
	Sagar Dawda
	Siddhant Jain
	Siddharth Pant
Team 2	Alex Issac
	Himalaya Ashish
	Karan Ingle
	Rahul Jain
	Rupali Hangekar
	Sankalp Saxena
	Shardul Nerlekar
	Vinod Boga
Team 3	Adeel Haidery
	Arun Kutty
	Darshin Doshi
	Mohammed Sunasra
	Nikhil Akki
	Nikhil Prasad
	Sanket Bhatt
Team 4	Ajaz Shaikh
	Melvin Jose
	Pawan Mathur
	Rohan Damodar

	Tanmay Sonar
	Tejesh Papineni
	Vighneshwar Jaiswal
Team 5	Ameya Dhale
	Ankan Roy
	Anujay Saraf
	Heramb Dharmadhikari
	Jeet Mukherjee
	Manish Kembral
	Nikita Antony
Team 6	Abhinav Anand
	Rajesh Upadhyay
	Saiprasad Balasubramanium
	Sandip Baradiya
	Shival Patel
	Utkarsh Garg
	Yuvraj Kale