Predicting high school scores, song popularity and start-up future using Machine Learning

Midhun Satheesan
School of Computing
National College of Ireland
Dublin, Ireland
x19146035@student.ncirl.ie

***Abstract*—This document proposes methods to analyse and quantify regression or classification on three different data sets using the Knowledge Discovery and Data mining (KDD) methodology.**

I. Motivation

A. The Spotify Hit Predictor

The music we listen to can be quantified using its tempo, energy, acoustics, amount of speech, utilisation of instruments etc. By studying these features in the songs of the past, it is possible to predict a potential hit or a flop. Spotify, the popular music streaming service provides an API which can be utilised to fetch the necessary information [1].

B. Startup Investments

According to Forbes, 9 in 10 start-ups fail. Many of the promising ones get acquired by one of the major players in the market. Several recent articles [2] suggest that the mammoth firms like Alphabet, Apple, Facebook, and Microsoft have become so much influential as to become the most valuable public companies in the world. Crunchbase holds an ever growing statistics of innovative companies around the world [3]. This study aims to predict the future of a start-up by learning the nature of funding it has obtained.

C. List of IQ Scores and high school grades

The longitudinal study funded by the National Institute on Aging (R01 AG009775; R01 AG033285) at Wisconsin provides the data about the life course of 10,317 men and women picked randomly. The subjects had taken Henmon Nelson test for IQ during the course of their education. Higher IQ does not always translate to better high school performance. This research aims at correlating the IQ scores to the students' high school ranks[4].

II. Research Question

A. The Spotify Hit Predictor

The objective is to predict a potential hit analysing the comprehensive list of songs from 1960 to 2019.

B. Startup Investments

The motive is to predict the future of an upcoming startup.

C. List of IQ Scores and high school grades

This research predicts the high school rank of students based on their IQ.

III. Initial Review

A. The Spotify Hit Predictor

Apart from the external factors like popularity of an artist, commercial influences etc. musical hits can be related to its audio quality too. The reference [5] points to a study utilising convolutional neural networks. The approach considers a regression problem and evaluations are done on data obtained from KKBOX Inc. - a music streaming service popular in the south east of Asia.

B. Startup Investments

Unfortunately,there are no certain methods for a venture capitalist to predict the success of a start-up. Earlier studies [6], to an extent, have tried to clarify this uncertainty by utilising data driven machine learning approach. This study try to extend the existing research by considering more dependant variables.

C. List of IQ Scores and high school grades

WLS have always been the subject of high quality studies based on people's personality, life style, health, and education. Also, the effects of peers, family, and IQ are all thoroughly correlated and studied as deciding factors of high school performance [7]. This study revisits the aspect of IQ as a deciding factor for high school performance.

IV. Data Sources

A. The Spotify Hit Predictor

The reference [1] lists the features of tracks fetched by Spotify's web API. The features include energy, key, loudness, modality, amount of spoken words, how acoustic the track is, liveliness, valence, tempo, duration, time signature, chorus hit and the number of sections in the song. The tracks are marked as a hit or a flop depending on certain criteria.

B. Startup Investments

The reference [3] lists the information about start up companies and investments. The source of the data is Crunchbase, which is a website that tracks investments and funding information of start ups. The data suggests if a company is still operating, closed down or acquired by another firm.

C. List of IQ Scores and high school grades

The Wisconsin Longitudinal Study (WLS) facilitates the study of the course of life of 10,317 men and women from late adolescence through 2008 [4]. This study on a glimpse of a data from the main WLS database captures the Henmon Nelson IQ test scores and their corresponding high school ranks. The Henmon Nelson tests of mental ability are intelligence measuring techniques used for children in 4 age groups - Grades 3-6, 6-9, 9-12, and college level.

V. Machine Learning Methods

A. Logistic Regression

Logistic regression is a method that can be used when the target is categorical. The target values in the Spotify predictor data are 0 (flop) and 1 (hit). This suggests that binary logistic regression can be used. The target values of Crunchbase data are operating, closed and acquired. This is a valid use case for the application of multinomial logistic regression.

B. Naive Bayes Classifier

Naive Bayes Classifier is a probabilistic classifier which is backed by the Bayes theorem of conditional probability. It assumes that all the independent variables contribute autonomously in the prediction of the response variable. It calls for the need for Gaussian naive Bayes since the predictors in the classification problems in hand are having continuous values.

C. Random Forests

A random forest consists of several individual decision trees working as an ensemble. The output is either the mode of the classes (classification) or mean prediction of the individual trees (regression).

D. Support Vector Machine

SVM is a non-probabilistic binary linear classifier which constructs a hyperplane or a set of hyperplanes in a high dimensional space. Support vectors are subsets of training points in the decision function. SVM is also a very memory efficient method for regression and outlier detection.

E. k-nearest neighbours algorithm (k-NN)

k-NN is a non-parametric method used for classification and regression.It assumes that similar entities exist in close proximity. The processing can get slow if the data is very large.The k-NN regression can be used to solve the regression problem of WLS data.

VI. Evaluation Methods

A. Confusion Matrix

It is a table which categorises testing data to true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).Following evaluation metrics are based on the Confusion Matrix.

1) Accuracy: Accuracy is the percentage of total items classified correctly.

2) Recall or Sensitivity: It is the number of items correctly identified as positive out of true positives(TP + FN)

3) Precision: It is the number of items correctly identified as positive out of total items identified as positive.

4) Specificity: It is the number of items correctly identified as negative out of total negatives. Specificity is the exact opposite of Recall.

B. R Squared

The coefficient of determination is the ratio between how good our model is vs how good is the naive mean model.The range of R squared is between negative infinity and 1. This metric can be used to evaluate the regression problem of WLS data.

C. Cohen's kappa

Kappa compares an observed accuracy with an expected accuracy (random chance). It is less misleading than other metrics like accuracy. The range of the values are between 0 and 1.

References

[1] F.Ansari, "The Spotify hit predictor dataset," 2020. [Online]. Available: https://www.kaggle.com/theoverman/the-spotify-hit-predictor-dataset [Accessed on: Feb. 18,2020].

[2] N. Patel, "90 percentage Of Startups Fail: Here's What You Need To Know About The 10 percentage," 2015.[Online]. Available: https://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/ [Accessed on: Mar. 10, 2020]

[3] M.Arindam, "Startup investments (Crunchbase)," 2017. [Online]. Available: https://www.kaggle.com/arindam235/startup-investments-crunchbase [Accessed on: Feb. 21,2020].

[4] Herd, Pamela, Deborah Carr, and Carol Roan. 2014. "Cohort Profile: Wisconsin Longitudinal Study (WLS)." International Journal of Epidemiology 43:34-41 PMCID: PMC3937969.

[5] L. Yang, S. Chou, J. Liu, Y. Yang and Y. Chen, "Revisiting the problem of audio-based hit song prediction using convolutional neural networks," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 621-625.

[6] J. Arroyo, F. Corea, G. Jimenez-Diaz and J. A. Recio-Garcia, "Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments," in IEEE Access, vol. 7, pp. 124233-124243, 2019.

[7] J Zax and D Rees, "IQ, Academic Performance, Environment, and Earnings," in Review of Economics and Statistics, vol. 84, pp. 600-616.