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

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

***Abstract*—No matter what culture we are from, music is extremely therapeutic that it underpins humanity. We live in a time where any music exists digitally and is available in the convenience of our personal abode. Curiosity of what factors sum up to the popularity of a song led to this study which uses song meta data from Spotify to classify songs as a hit or a flop. Decent levels of accuracy were obtained in this study. On a similar note, success of a start-up is judged as if it is still operating independently or has been acquired by a mammoth firm. Crunchbase has an enormous amount of data about start-ups and companies. This data is used for creating machine learning models to explore possibility of acquisitions based on funding scenario of the firm. Varios sampling techniques were utilized on unbalanced categorical data for creating the best models for prediction. On a third study in this paper, a cohort of over 10 thousand citizens from the city of Wisconsin are compared upon their highschool scores and Henmon Nelson IQ test scores in different levels of education. Supervised Machine learning methods are used to support the studies in this paper.**

***Index Terms*—music, start-up, acquisition, education, intelligence**

I. Introduction

A. The Music Hit Predictor

20 years into the millennium, physical forms of music sales have plummeted, but this does not mean that people have stopped listening to music. Thanks to online music streaming platforms, humans are listening as much as ever in their homes, offices, or any other location of their convenience utilising internet connectivity. These platforms are so vital to the music industry that the Recording Industry Association of America (RIAA) now considers music streaming into its Gold and Platinum album certifications [1]. This has a quintessential edge when it comes to analysing music as a digital representation. It is easier than ever to capture the meta features of a track to try to study if they contribute to the popularity and success of a song.

The music we listen to can be quantified using its tempo, energy, acoustics, amount of speech, utilisation of instruments, song time etc. Spotify - the Swedish music streaming service giant caches all these features of every song in the library.My intention is to create a machine learning model where these quantities are utilised to classify a song into a potential hit or a flop [2]. During this study, the condition for a track being ’flop’ is as follows:

- The track must not appear in the ’hit’ list of that decade.
- The track’s artist must not appear in the ’hit’ list of that decade.
- The track must belong to a genre that could be considered non-mainstream and / or avant-garde.
- The track’s genre must not have a song in the ’hit’ list.
- The track must have ’US’ as one of its markets.

B. Startup Investments

Young entrepreneurs find opportunities everywhere. The obvious first step to capitalise on that opportunity is to kick start a start-up. A start-up is a young venture aiming at solving a business problem where the solution is less obvious. Even though every entrepreneur crave for success, not everybody achieves it. The dream of every start-up is to either get acquired or to become public for the big money. According to Forbes, 9 in 10 start-ups fail [3]. Many of the promising ones get acquired by one of the major players in the market [4]. Several recent articles 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 [3].

The current ecosystem defines the success of a startup to two possibilities. Either to get acquired by a big player in the domain who have the necessary funding and skill set to fast track the development or to gain adequate funding to create their own legacy.Crunchbase is platform for finding business information regarding investments, funding, mergers and acquisition of private and public companies. Crunchbase holds an ever growing statistics of innovative companies around the world [5]. This study aims to predict the course of a start-up by learning the nature of funding it has obtained. The enlistment being studied has a huge list of company details along with their funding details and current status. The companies are either closed down, acquired by a big company or still running independently. Our focus is on the two positive paths: On companies which are still running or have been acquired.

C. List of IQ Scores and high school grades

Who wouldn’t love their child to be acing in school education ? It is believed that children with superior intelligence quotient (IQ) tend to learn more of what is taught in school than their peers with less IQ scores. This in turn converts to better grades in high schools. But, IQ is never a constant. It can vary according to the subject's experiences and course of life. IQ can be affected negatively in circumstances when pupil have extended breaks in studies due to lack of mind stimulating activities. However there can be a positive impact in IQ if the student stay in the course of learning [6]. 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 [7]. The subjects had taken Henmon Nelson test for IQ during the course of their education. 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.This study tries to explore the possibility that 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. this is in the belief that machine learning can aid in shedding light on further valuable insights on the correlation between the IQ of students in different stages in schooling and the high school performances.

The section II discusses few previous studies related to the analysis done in this paper. All the sections are divided into subsections based on which data it corresponds to. Section III contains all the details regarding the approach taken to create machine learning models required for analysis. The processes follow KDD methodology. Section IV evaluates and discusses the findings from the process done in section III. Section V conclude the studies by shedding light on planned future course of research.

II. Related work

A. The Spotify Hit Predictor

There have been a great number of efforts to learn what makes a song popular. Salganik, Dodds, and Watts undertook a popularity study that concentrated extensively on popularity's social impact.They find that the content of a song only partially determines whether a song is popular or not, and that social influence has an extremely important role to play [8].

Even though highly questionable, similarity to a previous hit song can attract the audience. J. Fan and M. Casey have attempted to study the regional differences in music traits in order to classify hit music [9]. This is under the presumption that people prefer music with features close to their culture.

A multimedia retrieval system (MRS) exist as an efficient mode of indexing, storing and handling the descriptive features of any form of media. The domain of multimedia retrieval is a highly researched area. The explosion of the quantity and diversity of media content have necessitated the evolution of every existing multimedia retrieving methodologies and algorithms. Due to the emergence of music streaming services like Spotify, Tidal and Lastfm music information retrieval (MIR) deserves attention. In [10], D. Martín-Gutiérrez et al have discussed an architecture for predicting song popularity with emphasis on both audio and text-based descriptors. This approach resulted in a better prediction rate than past studies. R. Gasser, L. Rossetto, and H. Schuldt introduces a multimedia retrieval stack named vitrivr which is the first system that seamlessly integrates support for four different types of media [11].

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 [12] 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

The course of a start-up is a highly uncertain one. Great ideas tend to fail due to lack of funding. A large crowd of entrepreneurs face difficulties just because the timing of their execution is not right. Any idea requires proper resources to be materialised. And proper resources do not come cheap. Venture capital (VC) industry provides investment opportunities in early-stage companies where volatility is high. Sadly, the solutions currently available to investors are not versatile enough to reduce risk and help them better handle uncertainty. The work by J. Arroyo et al. has shown that early-stage investors with no significant quantitative data or track record can decide on a potential investment in the baseline screening utilising a multi-class machine learning classifier that will help to improve an investor's success rate [13]. G.Xiang et al.have proposed to predict acquisitions considering technology news [14]. They have designed the study incorporating various other textual data features and credibility of the founder to explore start-up acquisitions. In a study by L. Y. Eugene and S. D. Yuan have tried to relate social relationships between investors and companies. Their take is that there is an accelerated chance of obtaining funding if there are healthy social media connections and relationships[15].

Crunchbase provides a large collection of data required for anything related to start-ups. Prediction of start-up success is a challenging study. I would like to revisit the study performed in [16] where various machine learning algorithms have been utilised for this attempt.

C. List of IQ Scores and high school grades

E. Winner in a case study suggests that the current educational system is less challenging for a superior, intellectually gifted child [16]. She compares America's curriculum with special educational programs available in European and East Asian countries. We often label students who crave for advanced learning exposure as gifted. Her point is that profoundly gifted children are very often under-challenged.

In a study [17], researchers tries to find the importance of intelligence quotient in the survival of a person. The research incorporates schooling performance into the study and suggests that people with higher consciousness have higher probability of surviving.

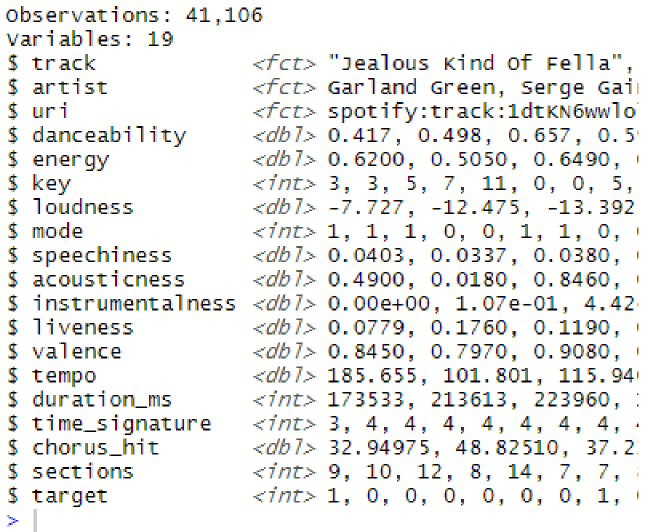
The Wisconsin Longitudinal Study have always been the subject of high quality studies based on people's personality, life style, health, and education. Also, the effects of peers, family, and IQ are all thoroughly correlated and studied as deciding factors of high school performance [7]. This study revisits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

III. Data Mining Methodology

All my study in this paper follows the Knowledge Discovery in Databases (KDD) approach to machine learning. The KDD process's unifying aim is to gather information from the data in large database context. A high level classification of the steps pertaining to the process include problem specification, data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. An additional step of sub sampling or super sampling is also required if the veracity of the model is found to be affected due to severe repercussions of imbalance of any sort.

A. The Music Hit Predictor

1) Pre-processing and transformations: The data consisted of music of 6 past decades divided into separate files of .csv formatting. All of them were imported to an R environment and combined into a single data frame. There were a total of 41,106 rows and 19 columns. The data points were checked for empty values. There were not any missing values. However there were non-numeric columns with meta data which were not of interest in this study.



```
Observations: 41,106
Variables: 19
$ track            <fct> "Jealous Kind of Fella",
$ artist           <fct> Garland Green, Serge Gai
$ uri              <fct> spotify:track:1dtKN6wwlo
$ danceability     <dbl> 0.417, 0.498, 0.657, 0.5
$ energy           <dbl> 0.6200, 0.5050, 0.6490,
$ key              <int> 3, 3, 5, 7, 11, 0, 0, 5,
$ loudness         <dbl> -7.727, -12.475, -13.392
$ mode             <int> 1, 1, 1, 0, 0, 1, 1, 0,
$ speechiness      <dbl> 0.0403, 0.0337, 0.0380,
$ acousticness     <dbl> 0.4900, 0.0180, 0.8460,
$ instrumentalness <dbl> 0.00e+00, 1.07e-01, 4.42
$ liveness         <dbl> 0.0779, 0.1760, 0.1190,
$ valence          <dbl> 0.8450, 0.7970, 0.9080,
$ tempo            <dbl> 185.655, 101.801, 115.94
$ duration_ms      <int> 173533, 213613, 223960,
$ time_signature   <int> 3, 4, 4, 4, 4, 4, 4, 4,
$ chorus_hit       <dbl> 32.94975, 48.82510, 37.2
$ sections         <int> 9, 10, 12, 8, 14, 7, 7,
$ target           <int> 1, 0, 0, 0, 0, 0, 0, 1,
> 
```

The columns track, artist and uri were removed. Now there are a total of 15 predictors that can be used to predict the target. It was noted that mode and target are listed as integers. They were converted to a nominal data type in R.

A principal component analysis is done to check if there are any columns contributing redundant information. No significant redundancy was found. The continuous valued predictors are then subjected to feature scaling to standardize the quantities. The data points were then divided into a training set and test set in the ratio 4:1.

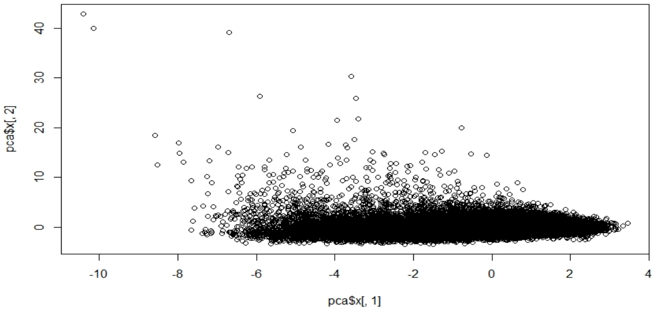


Fig. 1. PCA 1 vs PCA 2 bar plot

2) Prediction: To compare the qualities of algorithm multiple methods were used to create classification models to predict a potential hit music. The models were created utilising all the predictors. The methods used are Naïve Bayes classifier and logistic regression. Naïve Bayes expects the features leading to a model to be independent, which would not always be the case. When the dependencies become extreme the algorithm fails. Logistic regression is a linear method of classification, learning the likelihood of a sample belonging to a certain class. Logistic regression attempts to find the optimal boundary of decision which best separates the groups. When it comes to learning mechanism, Naive Bayes is a generative model and Logistic regression is a discriminative model. While trying to predict hits and flops in music logistic regression model was giving out higher accuracy rates than the Naïve Bayes model. I felt there is still scope for improvement. Hence random forest classification is implemented to create a model. The random forest algorithm is a decision tree based algorithm were a large number of uncorrelated trees work as an ensemble towards a single motive. A model created using random forest algorithm with 120 trees outperformed both Naïve Bayes and the logistic regression classifiers by a distance.

B. Startup Investments

1) Pre-processing: The data set from Crunchbase with the course of a start-up had 54,294 data points and 39 features in it. Many of the features were textual data or insignificant for the analysis. Hence, a new data object was created with specific columns - seed, venture, angel, debt financing, funding rounds count and 3 rounds of funding (rounds A,B and C). On checking for missing values (Fig.2) in the data set it was evident that there were rows without any data. after removing all the empty values the total number of data points reduced to 48,124.

The next course of action was to remove all the start-ups which did not belong to the categories - acquired or still operating. After removing them, the column which was textual was transformed to a nominal data type where 0 is assigned to acquired ones and 1 to still operating ones.Then, the numbers within each category were checked. 3,692 of the companies were acquired by a larger firm and 41,829 of them

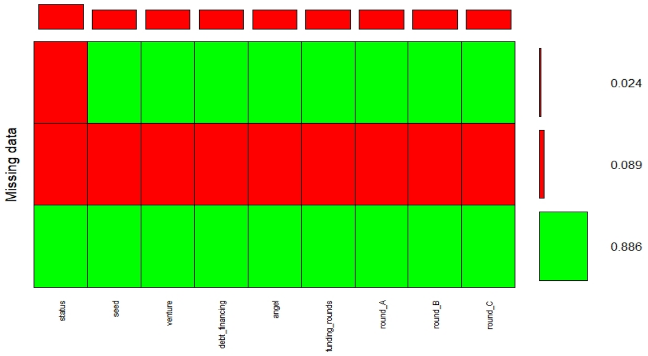


Fig. 2. Pattern of missing values

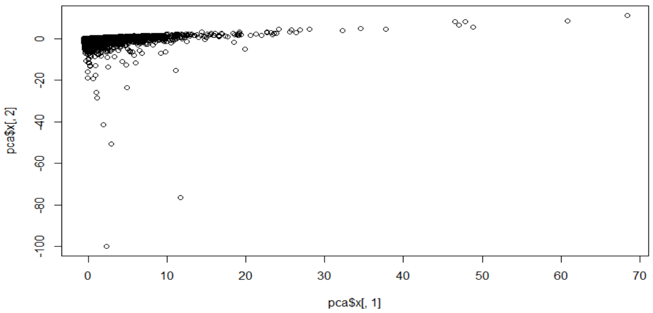
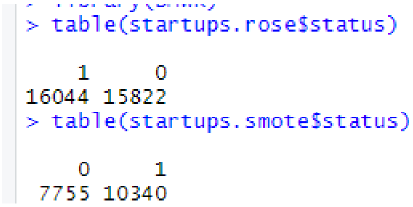


Fig. 3. PCA 1 vs PCA 2 bar plot

were still operating. With severely high numbers in a single category, the data set is deemed to be highly biased. Machine learning algorithms on classification data are designed for data with similar counts within categories. Classifiers created from dangerously biased training sets tend to predict the leading category more often. In this case, if such a classifier always predicts a start-up to be still operating, a very high accuracy will be obtained since 92 percent of the training data belonged to that particular category. A solution to this extremely common problem is sampling which is addressed in the following section.
Principal component analysis was done to check the variance the features offer.
Another interesting observation about the data set is that many chosen predictors are sparse. This study assumes that the value zero denotes no funding and considers it as a valid data.

2) Sampling to remove imbalance: Sampling are a wide range of techniques applied to the training data to ensure fair representation of every categories. It is ideally performed solely on training data after separating a part of main data set for testing the model. Keeping the test data free from synthesized data is important to ensure that evaluation of the model do not present misleading information. The start-up data is split into training set and a test set using a stratified sampling method to ensure the representation of both the categories in the test as well as training data. Two different sampling techniques were implemented to check the suitability of the resulting prediction model. Random over sampling examples (ROSE) is a smooth bootstrap based technique which generates a synthetic balanced sample of approximately equal size as the original data where the number of data points for both classes evenly present. Another method chosen is synthetic minority oversampling technique (SMOTE). C.Tantithamthavorn, A.E Hassan and K.Matsumoto suggests that SMOTE results in artificial data based on the feature space similarities from the minority modules [18]. Sample A was created using ROSE and sample B was created using SMOTE with nearest neighbour count of 10.



```
> table(startups.rose$status)

    1     0 
16044 15822 
> table(startups.smote$status)

    0     1 
 7755 10340 
```

Fig. 4. ROSE and SMOTE category population

3) Prediction: In order to try to predict the status of startups in test data, multiple models were created with sample A and sample B as the training set. The methods used were support vector machine classification (SVM) and decision tree algorithm. SVM is a non-probabilistic linear binary classifier modeling a hyper-plane or a collection of hyper-planes in a high-dimensional space. Support vectors in the decision function are subsets of training points. A decision tree is a tree in which each node represents an attribute, each connection is a decision and each leaf is an outcome. Additionally logistic regression is also applied. On sample A, The SVM classifier fared better over logistic regression classifier and decision tree classifier. The decision tree classifier had the worst performance.

C. List of IQ Scores and high school grades

This data was sourced from a long term study on a cohort of 10,417 people at the state of Wisconsin.

1) Pre-processing: Data was subjected to pre-processing and transformations in an R environment. According to the longitudinal study manual, -3 in the data represented lack of response. All the values less than zero was decided to be not inclusive in the study and rows with them were removed.

An alternate approach with imputation was considered. But it did not make sense as the response of an individual is assumed to be extremely unique. Hence it was decided to carry forward with the study with original and complete responses instead of synthesizing the data. After the cleaning, there were 6,044 subjects. On performing collinearity diagnostics, it was noted that the variables were highly correlated between each other. The model with severe collinearity can only be trusted for prediction. The role of each feature in the model cannot be analysed due to the dependencies between each other. The variance inflation factors were checked for each variable. It

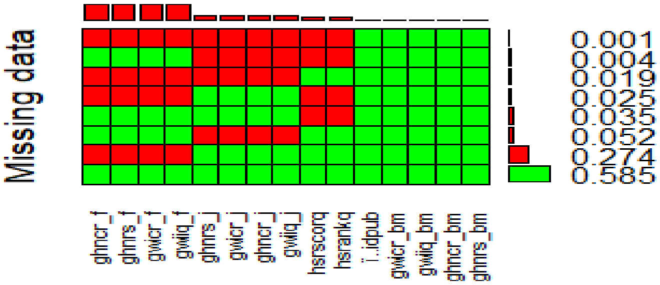


Fig. 5. Pattern of missing values

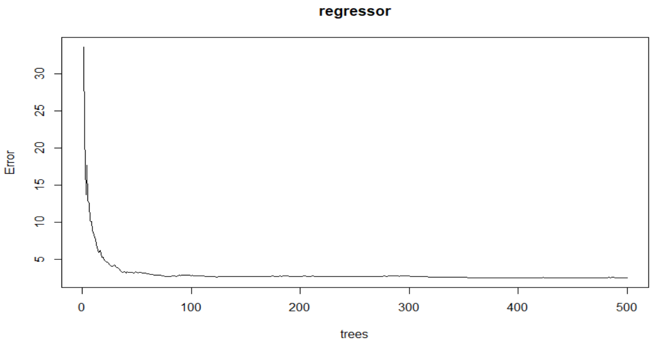


Fig. 6. Random Forest error plot

was found to be extremely high for most of the features. As multicollinearity induces redundancy in the data set it was necessary to proceed by performing feature reduction. Mean centering was applied to the variables as a way to reduce multicollinearity. This process had trivial impact which led to carry on with the original data.

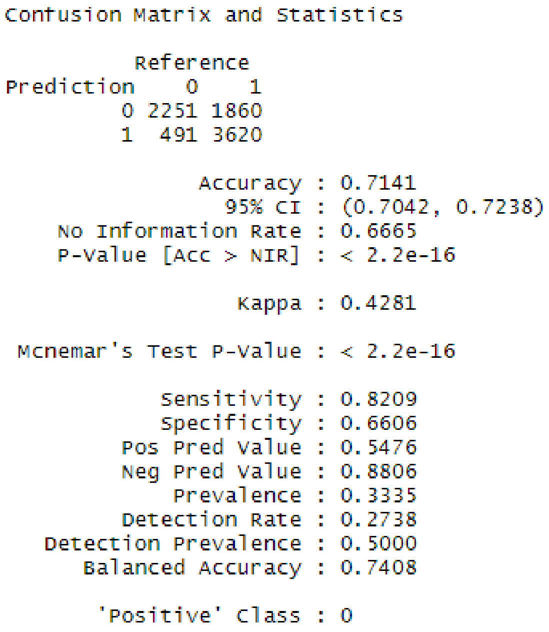
2) Creation of model: A regression model is created using multiple linear regression. It was expected that the model created would be insufficient in acknowledging the significance of the predictors. To segregate the variables backward elimination method was used. After a series of trial and error, it was noted that the only significant variables that contributed to the modeling of data are: IQ score mapped from raw freshman year Henmon-Nelson test, centile rank based on National test takers for Henmon-Nelson.

To support the result, a model was created using Random Forest regression. The ability of the model to predict values was satisfactory.

The error plot suggests that as the number of trees increased residuals decreased accordingly.

IV. Evaluation

Performance of the models created are evaluated using values derived from their corresponding confusion matrices such as accuracy, specificity, true positive rate, false positive rate, and precision. Receiver operating characteristic (ROC) curve was plotted to measure the diagnostic ability of the models.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2251 1860
         1  491 3620

               Accuracy : 0.7141
                 95% CI : (0.7042, 0.7238)
    No Information Rate : 0.6665
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4281

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8209
            Specificity : 0.6606
         Pos Pred Value : 0.5476
         Neg Pred Value : 0.8806
             Prevalence : 0.3335
         Detection Rate : 0.2738
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.7408

       'Positive' Class : 0
```

Fig. 7. Naive Bayes

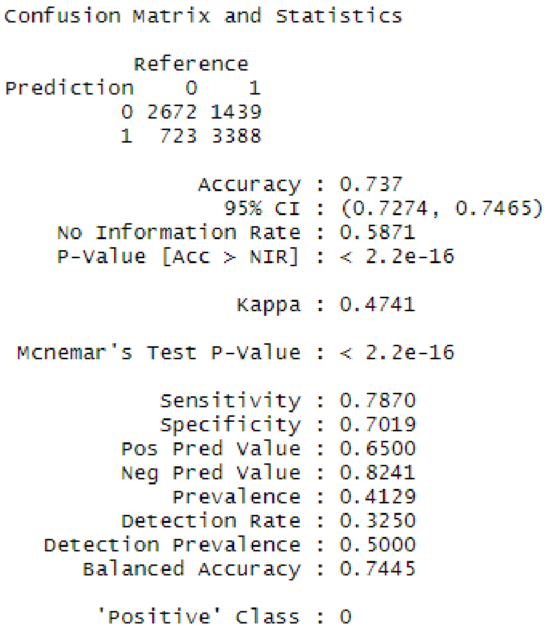
A. *The Music Hit Predictor*

Of the 3 classifiers explored to predict the success and failure of songs, The Random Forest classifier performed the best in giving accurate predictions with an accuracy of close to 79%. The Naive Bayes classifier did the worst in prediction. The Naive Bayes classifier gave an accuracy of 71%.

The true positive rate is considered to assert how many of the hit predictions were spot on in a model. Even though the accuracy was poor, he naive bayes model predicted better percentage of hits. This points to the poor prediction rate of flops by the naive bayes classifier. Random forest classifier had the better balance in categorising the inputs. The most optimal ROC curves from the models created are discussed in the following sections.

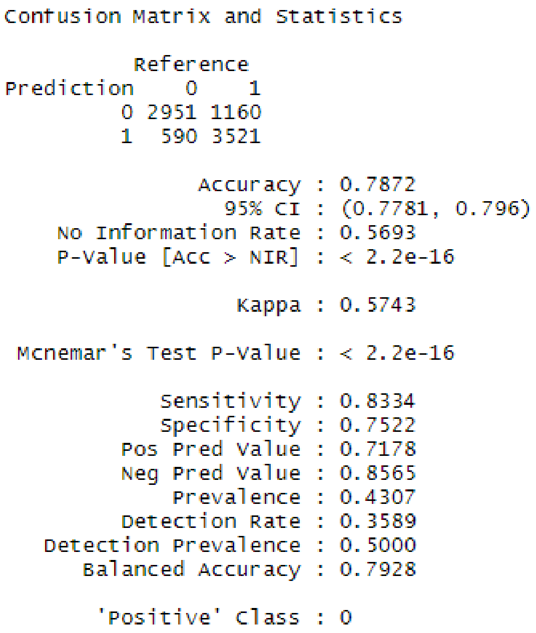
The Cohen's Kappa is a calculation of how well the classifier performed relative to how well it would actually perform by chance. According to standard scales, all 3 models had weak to moderate inter rater reliability. Random Forest model led with the best reliance with 0.57 kappa. The logistic regression model and the Naive Bayes classifier followed with kappa of 0.47 and 0.43 respectively.

The area under the Receiving Operator Characteristics(ROC) curve is an important measure of model performance. This value is called AUC. This graph is plotted as true positive rate against the false positive rate. The AUC for the random forest model was 0.787 and is considered optimum.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2672 1439
         1  723 3388

               Accuracy : 0.737
                 95% CI : (0.7274, 0.7465)
    No Information Rate : 0.5871
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4741

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7870
            Specificity : 0.7019
         Pos Pred Value : 0.6500
         Neg Pred Value : 0.8241
             Prevalence : 0.4129
         Detection Rate : 0.3250
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.7445

       'Positive' Class : 0
```

Fig. 8. Logistic regression



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2951 1160
         1  590 3521

               Accuracy : 0.7872
                 95% CI : (0.7781, 0.796)
    No Information Rate : 0.5693
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5743

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8334
            Specificity : 0.7522
         Pos Pred Value : 0.7178
         Neg Pred Value : 0.8565
             Prevalence : 0.4307
         Detection Rate : 0.3589
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.7928

       'Positive' Class : 0
```

Fig. 9. Random Forest

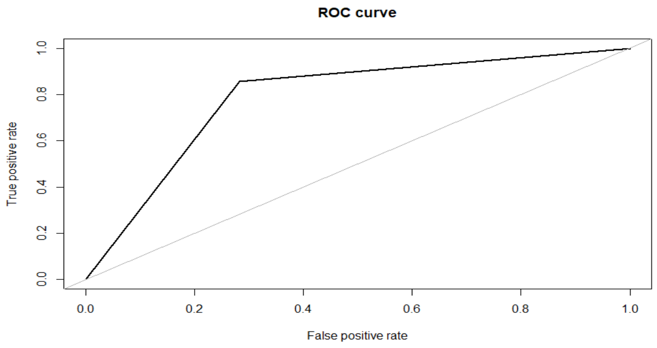
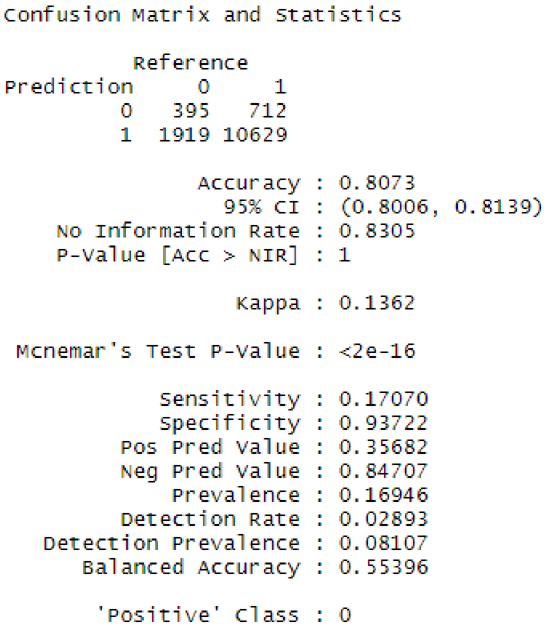


Fig. 10. ROC of Random Forest model



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   395   712
         1  1919 10629

               Accuracy : 0.8073
                 95% CI : (0.8006, 0.8139)
    No Information Rate : 0.8305
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1362

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.17070
            Specificity : 0.93722
         Pos Pred Value : 0.35682
         Neg Pred Value : 0.84707
             Prevalence : 0.16946
         Detection Rate : 0.02893
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.55396

       'Positive' Class : 0
```

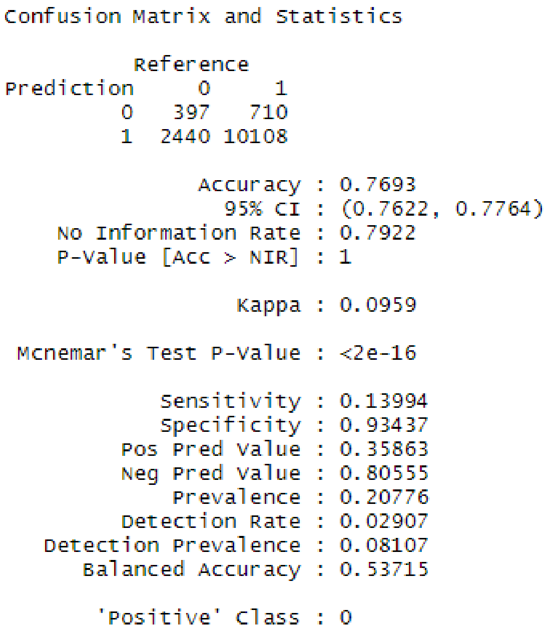
Fig. 11. ROSE Decision tree algorithm

B. *Startup Investments*

The 3 classifiers decision tree model, logistic regression model, and SVM model were created using 2 different sampling techniques.The data in its raw for had a severe class imbalance and always resulted in extreme accuracy of over 90%. This owed to the fact that the model learned to predict the major class more and played it safe by giving out the major class as response to new inputs often. This basic modelling had dangerously poor reliability.

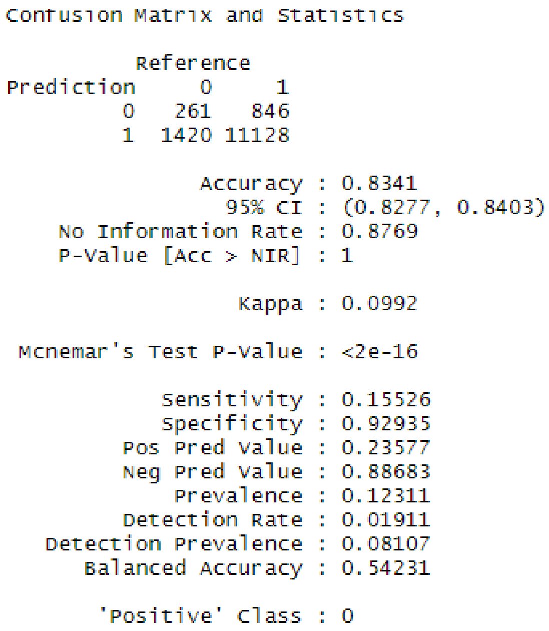
Hence re-modeling was done using 2 efficient algorithms. Training set of data were infused with synthesised data as an attempt to achieve better reliability. The sample A which utilised ROSE had higher number of training data points. Sample B relied on under-sampling using SMOTE.

1) ROSE Sampling: The SVM model gave better accuracy of 83% over the logistic regression model and the decision tree based model. The tree based modelling failed the worst



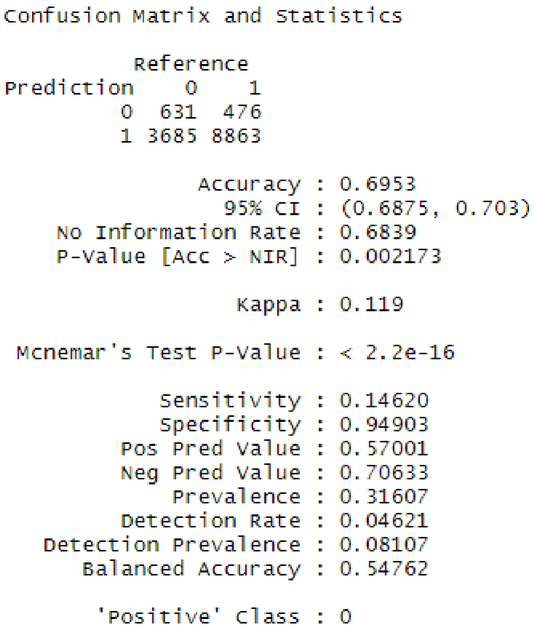
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   397   710
         1  2440 10108

               Accuracy : 0.7693
                 95% CI : (0.7622, 0.7764)
    No Information Rate : 0.7922
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0959

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.13994
            Specificity : 0.93437
         Pos Pred Value : 0.35863
         Neg Pred Value : 0.80555
             Prevalence : 0.20776
         Detection Rate : 0.02907
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.53715

       'Positive' Class : 0
```

Fig. 12. ROSE and Logistic Regression



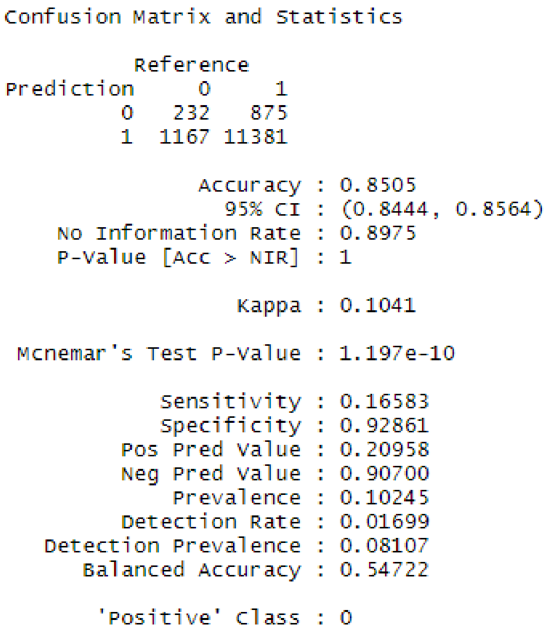
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   261   846
         1  1420 11128

               Accuracy : 0.8341
                 95% CI : (0.8277, 0.8403)
    No Information Rate : 0.8769
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0992

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.15526
            Specificity : 0.92935
         Pos Pred Value : 0.23577
         Neg Pred Value : 0.88683
             Prevalence : 0.12311
         Detection Rate : 0.01911
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.54231

       'Positive' Class : 0
```

Fig. 13. Confusion Matrix - ROSE and SVM



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  631  476
         1 3685 8863

               Accuracy : 0.6953
                 95% CI : (0.6875, 0.703)
    No Information Rate : 0.6839
    P-Value [Acc > NIR] : 0.002173

                  Kappa : 0.119

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.14620
            Specificity : 0.94903
         Pos Pred Value : 0.57001
         Neg Pred Value : 0.70633
             Prevalence : 0.31607
         Detection Rate : 0.04621
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.54762

       'Positive' Class : 0
```

Fig. 14. Confusion Matrix - SMOTE and Decision Tree Algorithm



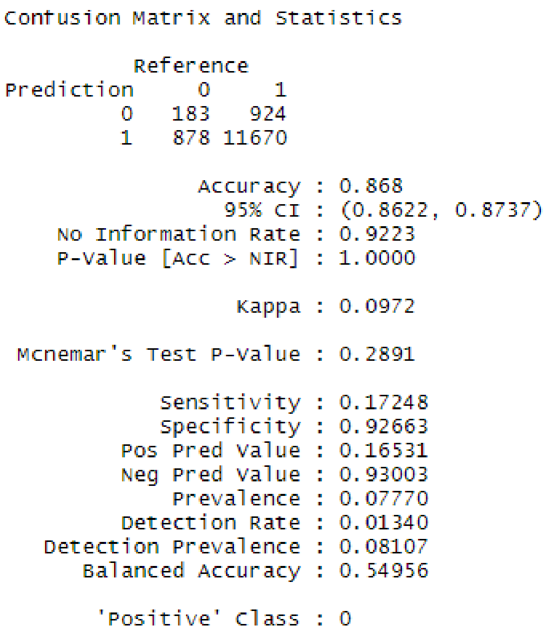
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   232   875
         1  1167 11381

               Accuracy : 0.8505
                 95% CI : (0.8444, 0.8564)
    No Information Rate : 0.8975
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1041

 Mcnemar's Test P-Value : 1.197e-10

            Sensitivity : 0.16583
            Specificity : 0.92861
         Pos Pred Value : 0.20958
         Neg Pred Value : 0.90700
             Prevalence : 0.10245
         Detection Rate : 0.01699
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.54722

       'Positive' Class : 0
```

Fig. 15. Confusion Matrix - SMOTE and Logistic Regression

in accuracy. The Cohen's kappa inter-rater reliability values were poor in all the models. This might have owed to the presence of too much synthesised as well as sparse data in the learning data set. The decision tree model had the highest kappa coefficient of reliance among all the 6 models explored in this analysis. The logistic regression model had the worst.

2) SMOTE Sampling: The training data in sample B was not as large as sample B. Once again the SVM model proved to be efficient in getting the result right. The test data consisted of more data from the major class. So it is meaningless to consider the true positive rate of prediction.

The area under the ROC for the models were around 0.60. The 3 funding rounds A, B and C were found to be features contributing most to the model along with the seed funding.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   183   924
         1   878 11670

               Accuracy : 0.868
                 95% CI : (0.8622, 0.8737)
    No Information Rate : 0.9223
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.0972

 Mcnemar's Test P-Value : 0.2891

            Sensitivity : 0.17248
            Specificity : 0.92663
         Pos Pred Value : 0.16531
         Neg Pred Value : 0.93003
             Prevalence : 0.07770
         Detection Rate : 0.01340
   Detection Prevalence : 0.08107
      Balanced Accuracy : 0.54956

       'Positive' Class : 0
```

Fig. 16. Confusion Matrix - SMOTE and SVM

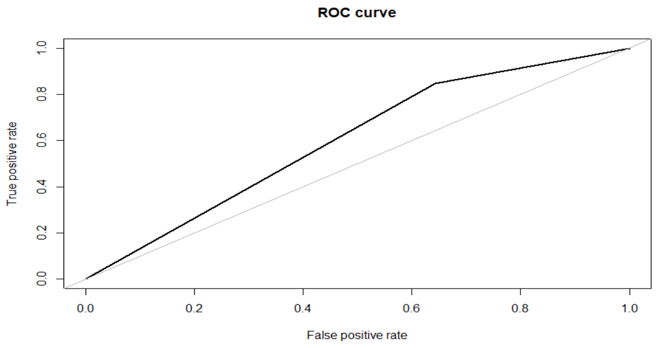


Fig. 17. ROC of model

C. List of IQ Scores and high school grades

On evaluation it was noted that the only features that effected the highschool grades were IQ score mapped from raw freshman year Henmon-Nelson test and centile rank based on National test takers for Henmon-Nelson. The model [Fig. 18] predicted values with an accuracy of 91%

V. Conclusions and Future Work

A. The Music Hit Predictor

The primary objective of the study was to predict a future hit music based on the features such as danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness etc. using machine learning algorithm and to decide which algorithm suited the purpose. Logistic Regression, Naive Bayes and Random Forest algorithms were used to train a model and were tested against new data. Evaluations on all the models proved that, the Random Forest model responded the best with getting right 78% of the predictions right.

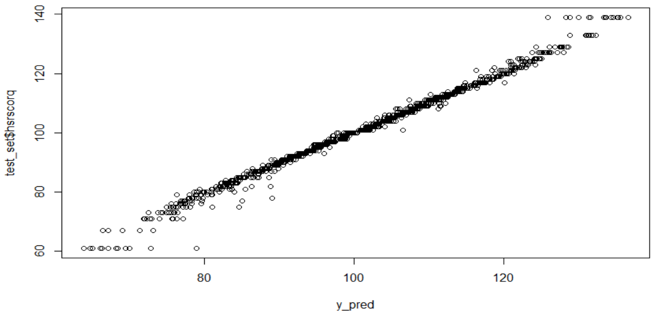


Fig. 18. Predicted vs actual

Even though few of the quantitative measures of what constitute a song are studied, an important aspect that has been missed out from the study is the lyrics of the song. It would be interesting to extend the study to this aspect.

B. Startup Investments

The objective of the study was to create a model to predict start-up acquisitions in the industry and to explore what factors affect the acquisitions. The data had to undergo intense processing to get ready for analysis due to severe categorical imbalance. Several approaches were taken using 2 different sampling methods: SMOTE and ROSE. For model creation, decision tree algorithm, SVM and logistic regression were used. The results varied significantly with various choice of algorithms and sampling techniques. The decision tree algorithm with SMOTE sampling was capable of predicting the highest number of acquisitions correctly. But it failed in properly getting the other category right.SVM with SMOTE sampling was the leader in that aspect. It would be interesting to extend the study with more valid data in the minor category instead of infusing synthetic sampling. This would enable to create models with much more realistic nature. The presence of sparse values in the data demands a study by transforming the factors into categorical data based on thresholds. Thi might also result in new intuitions about the acquisitions.

C. List of IQ Scores and high school grades

The objective was to know if the Henmon Nelson IQ scores of students in various stages of schooling can be related to the final high school score. Multiple linear regression with backward step-wise method was performed to reduce dimensionality and redundancy among the features. The study could find that the IQ of a student during raw freshman year can be associated with the final high school scores.
Further the study can be extended to relate IQ to various aspects in life such as mortality in a specific age.

References

[1] "RIAA DEBUTS ALBUM AWARD WITH STREAMS," RIAA, Feb. 01, 2016. https://www.riaa.com/riaa-debuts-album-award-streams/ (accessed April 02, 2020).

[2] 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].

[3] 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]

[4] A. Riani, “How To Make Your Early-Stage Startup Valuable To Acquirers,” Forbes. https://www.forbes.com/sites/abdoriani/2020/03/06/how-to-make-your-early-stage-startup-valuable-to-acquirers/ (accessed May 02, 2020).

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

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

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

[8] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market,” Science, vol. 311, no. 5762, p. 854, Feb. 2006, doi: 10.1126/science.1121066.

[9] J. Fan and M. Casey, “Study of Chinese and UK hit songs prediction,” in Proc. Int. Symp. Comput. Music Multidisciplinary Res., 2013, pp. 640–652.

[10] D. Martín-Gutiérrez, G. Hernández Peñaloza, A. Belmonte-Hernández, and F. Álvarez García, “A Multimodal End-to-End Deep Learning Architecture for Music Popularity Prediction,” IEEE Access, vol. 8, pp. 39361–39374, 2020, doi: 10.1109/ACCESS.2020.2976033.

[11] R. Gasser, L. Rossetto, and H. Schuldt, “Towards an all-purpose contentbased multimedia information retrieval system,” 2019, arXiv:1902.03878. [Online]. Available: http://arxiv.org/abs/1902.03878

[12] 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.

[13] 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.

[14] G. Xiang, Z. Zheng, M. Wen, J. Hong, C. Rose, and C. Liu, “A Supervised Approach to Predict Company Acquisition With Factual and Topic Features Using Profiles and News Articles on TechCrunch,” p. 8.

[15] L. Y. Eugene and S. D. Yuan, ”Where’s the Money? The Social Behavior of Investors in Facebook’s Small World,” 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 158-162.

[16] F. R. da Silva Ribeiro Bento, “Predicting start-up success with machine learning,” M.S. thesis, Dept. Inf. Manage., Universidade Nova do Lisboa,Lisbon, Portugal, 2018

[17] E. Winner, “Exceptionally High Intelligence and Schooling,” American Psychologist, p. 12, 1997.

[18] Hauser, R. M. and A. Palloni. 2011. ”Adolescent IQ and Survival in the Wisconsin Longitudinal Study.” The Journals of Gerontology.Series B, Psychological Sciences and Social Sciences 66 Suppl 1: i91-101. www.scopus.com.

[19] C. Tantithamthavorn, A. E. Hassan, and K. Matsumoto, “The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models,” arXiv:1801.10269 [cs], Jan. 2018, Accessed: Apr. 27, 2020. [Online]. Available: http://arxiv.org/abs/1801.10269.