Major project 02_22.pdf
May 31, 2021
11363 words / 58809 characters

Major project 02_22.pdf

Sources Overview

3%

OVERALL SIMILARITY

1	machinelearningmastery.com INTERNET	<1%
2	docshare.tips INTERNET	<1%
3	www.slideshare.net INTERNET	<1%
4	link.springer.com INTERNET	<1%
5	www.tandfonline.com INTERNET	<1%
6	Kangbok Lee, Sunghoon Joo, Hyeoncheol Baik, Sumin Han, Joonhwan In. "Unbalanced data, type II error, and nonlinearity in predicting ... CROSSREF	<1%
7	www.issc.org INTERNET	<1%
8	IIT Delhi on 2018-06-06 SUBMITTED WORKS	<1%
9	towardsdatascience.com INTERNET	<1%
10	www.gwu.edu INTERNET	<1%
11	gcoej.ac.in INTERNET	<1%
12	hdl.handle.net INTERNET	<1%
13	Amity University on 2016-10-12 SUBMITTED WORKS	<1%
14	University of Lancaster on 2019-09-27 SUBMITTED WORKS	<1%
15	Xin-Li Yang, David Lo, Xin Xia, Qiao Huang et al. "High-Impact Bug Report Identification with Imbalanced Learning Strategies", 'Springer ... INTERNET	<1%
16	journals.sagepub.com INTERNET	<1%
17	popups.uliege.be INTERNET	<1%

18	rohit521989.blogspot.com INTERNET	<1%
19	www.airccse.org INTERNET	<1%
20	Ning Wang, Senyao Zhao, Shaoze Cui, Weiguo Fan. "A hybrid ensemble learning method for the identification of gang-related arson cas... CROSSREF	<1%
21	University of Malta on 2018-09-29 SUBMITTED WORKS	<1%
22	University of Reading on 2019-09-15 SUBMITTED WORKS	<1%
23	epub.ub.uni-muenchen.de INTERNET	<1%
24	zombiedoc.com INTERNET	<1%

Excluded search repositories:

- None

Excluded from Similarity Report:

- Bibliography
- Quotes
- Small Matches (less than 10 words).

Excluded sources:

- None

Major Project Report on

Analysis and Prediction of Mergers & Acquisitions

Submitted By:

Anureet Bansal (2K19/BMBA/02)

Kunal Rao (2K19/BMBA/22)

Under the Guidance of:

Dr. Gaganmeet Kaur Awal

Assistant Professor

UNIVERSITY SCHOOL OF MANAGEMENT & ENTREPRENEURSHIP

Delhi Technological University

Bawana Road, Delhi - 110042

May 2021

CERTIFICATE

This is to certify that Anureet Bansal (2K19/BMBA/02) and Kunal Rao (2K19/BMBA/22) are bona fide students of University School of Management and Entrepreneurship, Delhi, and have successfully completed the project work as prescribed by the Delhi Technological University in the partial fulfillment of the requirement of Master Of Business Administration (MBA), Business Analytics Program for the academic year 2019-2021.

The Project Work titled "Analysis and Prediction of Mergers & Acquisitions".

Project Guide

Dr. Gaganmeet Kaur Awal

Assistant Professor

DECLARATION

The work embodied in this report entitled "M&A prediction and its Analysis" submitted by us to the Delhi Technological University, in partial fulfillment of the requirement for the award of the degree of Master of Business Administration (MBA), Business Analytics under the guidance of Dr. Gaganmeet Kaur Awal, is our original work and the conclusions drawn therein are based on the material collected by ourselves.

The work submitted is original and has not been submitted earlier to any institute or university for the award of any degree or diploma. We shall be responsible for any unpleasant moment/situation.

Place: New Delhi

Date: 31th May 2021

Anureet Bansal (2K19/BMBA/02)

Kunal Rao (2K19/BMBA/22)

ACKNOWLEDGMENT

We would like to express our gratitude towards our faculties and family who gave us an opportunity to learn and succeed in our lives. We thank our colleagues and fellow research scholars for their constant support during the course of this project. We express a special vote of thanks to our mentor Dr. Gaganmeet Kaur Awal for her guidance. Our thanks and appreciation go to the entire USME, Delhi Technological University family who taught us the concepts that were beneficial during the experiments conducted for this project.

We finally extend our warm thanks to HOD sir for giving us a platform to present our work towards esteemed and renowned faculty members for the fulfilment of the requirement of this MBA course.

ABSTRACT

Sustaining in a highly competitive market is difficult and a challenge that the company looks to overcome by providing the best to customers than the existing options available. To diversify their operations and attain the position of the market leader, most companies opt for Mergers or acquisitions for having power or success in this changing environment. The black swan events have resulted in many Mergers, Acquisitions, and takeovers due to firms not able to adapt to changing times and the constantly innovating firms acquiring such targets which would benefit them in the longer term.

In this project, we aim to study and explore the multi-class prediction problem of identifying the status of the company whether it should opt for mergers, acquisitions, IPO, or continue in operating mode. Firstly, we provide exploratory data analysis using popular data visualization tools to gain useful insights from Crunchbase and WorldBank datasets. Secondly, we propose a novel method to address the prediction problem to identify the status of the company using machine learning techniques. We have employed various under-sampling methods to deal with the problem of imbalance in the dataset. Also, we incorporate the additional factors like macroeconomic variables and Intellectual property rights of the home country which are considered useful from an M & A perspective. We have performed experiments to determine the best-performing model among machine learning techniques like Logistic Regression, K-nearest neighbor, Random Forest, and XG Boost and compare the results with the baseline modeling using appropriate evaluation metrics. The Edited Nearest Neighbour under-sampling technique presents the best results using K-nearest neighbor and closely followed by XGBoost Classifier Model. Experimental results demonstrate that the proposed approach outperforms the existing methodology adopted by the researchers in the past.

Keywords: Machine Learning, Mergers, Acquisitions and imbalance learning, Edited Nearest Neighbour, Extreme Boosting Algorithm, Exploratory Data Analysis, K-nearest neighbors

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1. M&A	Mergers and Acquisitions
2. ML	Machine Learning
3. WB	World Bank
4. IP	Intellectual Property
5. ENN	Edited Nearest Neighbour
6. EDA	Exploratory Data Analysis
7. CNN	Condensed Nearest Neighbor
8. NCR	Neighbourhood Cleaning Rule
9. OSS	One Sided Selection

CHAPTER 1

INTRODUCTION

In an era where the world is becoming highly competitive, it becomes difficult to sustain. With this increasing competitiveness, the number of businesses in a particular domain or field also keeps on increasing. Every new company aims to provide the best to customers than the existing options available. In a need to have the most power or succeed, most companies opt for Mergers or acquisitions. The term M&A which is a common abbreviation for Mergers and acquisitions is used to describe the consolidation of two firms through financial transactions. Mergers are basically when two companies combine to form one big firm in such a way that one company ceases to exist and Acquisitions are when one company acquires a major stake of the other company by purchasing shares or acquiring the assets. There can be multiple reasons why a company opts for either mergers or acquisitions. Few reasons can be:

- Obliterate competition: One of the major reasons for M&A is to eliminate competition. Most of the time a big firm acquires or merges with another firm to stay ahead in the competition. With the help of this, they also achieve a higher market share.
- Diversification of business: When a firm wants to introduce new products or diversify its business it can opt for M&A. By this one can incorporate an established product or company and increase its overall profitability.
- Enhance capabilities of the firm: One of the major benefits is that firm's capabilities in research and development, economies of scale, and manufacturing system are enhanced.
- Tax Benefits: Many firms opt for M&A to gain tax benefits. If any company operates where tax is high, it can be merged with another company where tax rates are low. Also if any company has huge table profits it can be merged with a company of tax losses which in turn gets balanced.

There can be multiple reasons for M&A to occur but the end goal of each reason is the betterment of a firm.

Structure of mergers: Mergers can be of various types according to the relationship between the two parties involved. Few types of mergers are stated below:

- Horizontal merger: When the two firms are competitors and have the same product line
- Vertical Merger: In this, a company merges with its customer or a supplier with the company like a tomato seller merges with a ketchup company.
- Congeneric Merger: This occurs when two companies serve in different ways but to the same consumer base. For example: TV manufacturers and Netflix.
- Market extension merger: In this, the two firms sell the same product but in different markets. This takes place to increase reachability.
- Conglomerate ration: It occurs when two firms with no common business merge.

Types of M&A: Based on underlying transactions we can categorize mergers and acquisitions. In the case of mergers, two companies combine and shareholders approve the deal. In the case of acquisitions, the acquiring company takes the major stake of the other company but the structure of the firm remains the same whereas in the case of consolidations, a new firm is formed and old structures are abandoned. In asset acquisitions, the assets of a firm are acquired with the permission of its stakeholders, and last but not the least, in management acquisition or Management led buyout controlling stake is taken off a company.

Benefits: M&A is an important process that helps firms to increase their value or grow. This is achieved by acquiring valuable assets or intellectual property, new technologies, staff with useful skills and knowledge, increasing their consumer base, and attaining economies of scale.

Risk: Even though the objective of M&A is an advancement of any firm but few risks are also associated with it. There can be conflicts in the organization due to different company cultures. There are also chances that assets are less valuable as evaluated initially or the M&A can be a little expensive as thought. After the M&A, there's a possibility that resources are used for managing mergers and employees are reluctant to join the new organization.

13 Initial Public Offering (IPO): Initial Public Offering is that stage in which a private company plans to be listed on an exchange and goes public by selling its stocks to the general public. In the secondary market, these shares are further sold by investors So

it can be said that if a firm is going for an IPO, then the firm is operating in a good space and its valuation would be better and a good target for an M&A event.

1.1 Dataset Repositories

The dataset used for this project belongs to the Crunchbase organization and the WorldBank data repository.

Crunchbase is a data providing company which was founded in 2007. It provides business information for public companies as well as private companies. Besides the Mergers and acquisitions data that was taken for this project, it also provides data about industry trends, founding members, investments, and funding information of companies. The site obtains its data from four sources, one is through its venture program, the second is from machine learning, the third is from its community and the last is through its in-house data team.

WorldBank is an international institution created in 1944 along with the International Monetary Fund (IMF) at the Bretton Woods Conference. It consists of the International Bank for Reconstruction and Development (IBRD), International Development Association (IDA), International Finance Corporation (IFC), Multilateral Investment Guarantee Agency (MIGA), International Centre for Settlement of Investment Disputes (ICSID).

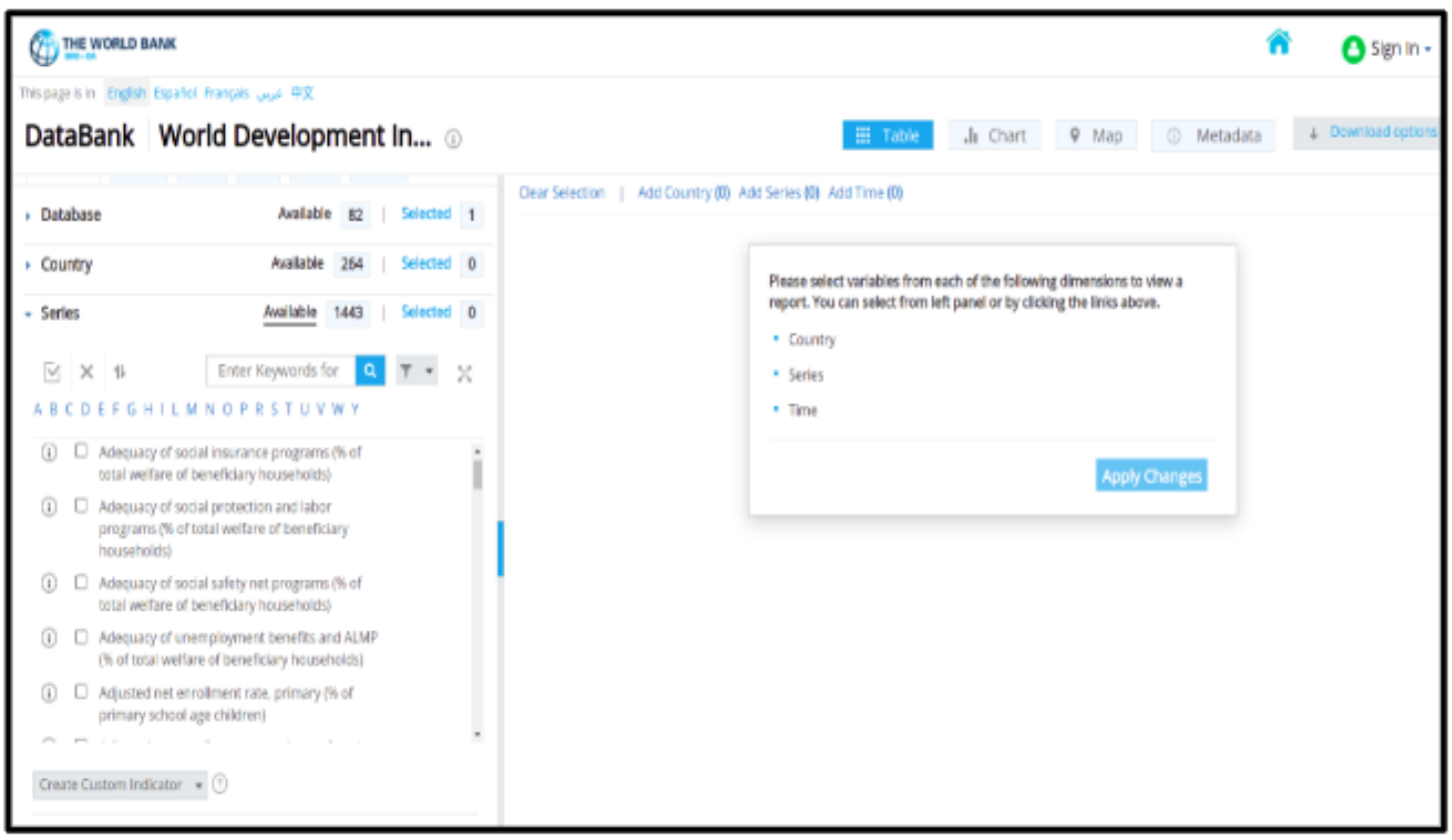

Fig 1.1: World Bank Open Source Dataset View

The World Bank Open Data gives us access to download the dataset related to Commodity Market, Climate Change, World Development Indicators, Gender Statistics, Health Nutrition, MDG, SDG, Poverty and Equity, and others as shown in Fig 1.1.

Further, the dataset for the GDP, PPP, Literacy, Employment, interest rates, patent and trademark, and other indicators can be accessed and for the last 50 years can be used as a period for basing an analysis. All 189 member countries have a dataset for all the indicators and that can be downloadable in the text, csv, or excel format.

1.2 Current Industry Scenario

The 1990 post-economic reforms and LPG (Liberalisation, Privatisation, and Globalisation) aided the Indian corporate culture and globally provided a huge market for big players. As the global players increase in the Indian market the indigenous players face stiff competition to compete locally and globally with them. The M&A events and target prediction are some of the most challenging events. The past 2 decades have seen many top firms being acquired.

Vodafone AirTouch acquired Mannesmann AG, a wireless pharmaceutical company for $ 202,785 million in 1999. AT&T acquired Time Warner, a media company for $ 88,400 million. Walt Disney Company acquired 21st Century Fox in 2018, Microsoft acquired LinkedIn in 2016. Dell agreed to an all buy-out by its founder for $13.65 per share in cash in 2013, The Vodafone and Idea Cellular merger in Indian Operations in 2017, S&P Global acquired IHS Markit in 2020 for a deal of $44 billion. Salesforce acquired Slack Technologies in 2020 and AstraZeneca, a leading pharmaceutical firm acquired an American pharmaceutical company Alexion for $39 billion in 2020.

Yet many deals like Pfizer acquiring AstraZeneca in 2010-2020 did not take place due to undervaluation and similarly other deals due to multiple reasons the M&A events could not take place.

So this valuation and target for M&A is an interesting study and to study the factors affecting these events and the industry profile gives us an idea of the deals that have surfaced in the past few decades.

1.3 Objective of the study

In this project we have divided our objectives into two parts. One is objectives for exploratory data analysis and the other one is objectives related to data modeling. The objectives for the study are listed below:

- **Exploratory data analysis**

1. Companies from which sector has been acquired the most and by which sector
2. Correlation between various variables to find the relationship between them.
3. To know the top countries for each status type.
4. To find the sector with the highest fundings.
5. To find out the top M&A valuation deal company-wise.

- **Data Modelling**

  Predicting whether the company should opt for M&A (acquired or closed), or they should offer IPO to raise money and continue working or they should keep on operating as usual.

1.4 Organization of the Report

The remainder of this report is organized as follows. Section 2 describes the Literature review that is the previous work done in the same domain by other researchers. Section 3 presents the research methodology adopted to conduct our study. Section 4 presents the experimental results and the dataset description and steps of the data preprocessing of our study. Section 5 discusses our findings and some recommendations as per our observations and finally, Section 6 concludes with the limitations of the study conducted. At the end of the report, references from which we referred and our plagiarism report has been provided.

CHAPTER 2

LITERATURE REVIEW

Machine learning tools are utilized in finance, investment, and valuation domain areas to predict company success or failure. During the previous research analysis, it was found out that using the Crunchbase dataset provides an optimal learning source to study the company analysis and prediction of company fortunes in the future. Some of the notable works published in the literature are mentioned below.

The Literature Survey consists of the following sections. Section 2.1 has a detailed review of the past research conducted on the Crunchbase dataset. Section 2.2 covers review into past research in the M&A events prediction and the algorithms being deployed by them. Section 2.3 covers description of Machine Learning used for M&A Prediction and Section 2.4 presents the Imbalance learning. Section 2.5 presents discussion about the under-sampling techniques explanation, Section 2.6 discusses challenges and limitations of the previous works and Section 2.7 presents our contributions.

2.1. Literature Work related to Crunchbase Dataset

Many researchers have found the Crunchbase data to be useful for analyzing the startup environment and drawing out interesting insights. Färber et al. (2018) proposed a model built on the Crunchbase dataset which crawls Resource Description Framework (RDF) data based on Linked Data API to build customized knowledge graphs. Liang and Yuan (2016) explored the Crunchbase dataset to build a social network of company profiles. They studied the nature of investing and explored investors and companies decisions to find interesting insight through social graphs. The link or relationship is formulated to predict whether investors create links with companies in the social graph. They used techniques like common neighbors, shortest path, Jaccard Coefficient to study insights into social networks.

Nathan et al. (2017) analyzed the Crunchbase data at the organizational level to fill up the gaps for the United States, Canada, and UK. Xiang et al. (2012) harnessed factual and topic features using the user profiles from the TechCrunch data repository to predict the acquisition of the company. Batista and Carvalho (2015) proposed the

Fuzzy fingerprint technique for developing a predictive model to predict the categories based on text classification models. Their model outperformed the other popular machine learning models like K-nearest neighbors, Naive Bayes applied in literature before.

2.2. Algorithms for M&A events prediction

Yang et al. (2014) proposed a new factor to predict M&A. 43 technological indicators were selected from the patent documents of the companies. These companies were acquired between the time period of January 1997 to May 2008 and were based in Japan and Taiwan. After collecting technological indicators along with technological profiles of both bidder and candidate target company, an ensemble learning model is applied. Their result shows that technological indicators improve the prediction of M&A than those attained using financial variables. Also, ensemble learning performed better as compared to a single learner in terms of accuracy and F1 score.

Lee et al. (2020) proposed a new approach in forecasting M&A failure or success. This approach resolves the three major issues associated with the traditional method of forecasting in which the first problem is that cases of failure of M&A are generally less which makes data imbalanced, the second issue is of Type-II error which occurs by misclassifying cases of failure as success and the third problem is of non-linear nature of data used for prediction. In this model, a neural network with a generalized logit activation function has been used to resolve the issue of imbalanced data and a cost-sensitive function is employed to handle the issue of misclassification. The dataset consists of M&A deals between the period 2009-2015. The new approach performed better than benchmark models applied in this research paper. The performance of the classification model was measured using multiple accuracy rate measures.

Liu et al. (2011) presented a two-stage multi-kernel algorithm for predicting the price of a candidate that is the target company in M&A. This method combines the advantages of various hyperparameter settings by using multiple kernel SVMR and showed better performance as compared to other models. The performance of the model was determined using the error rate. These results showed that the price predicted can be used for M&A decisions.

2.3. Description of Machine Learning used for M&A Prediction

In this section, we will give a brief explanation of machine learning models deployed on the dataset. In this project, the baseline model is the original dataset without learning from the imbalance learning.

Logistic Regression: Logistic regression is a binary classificathe tion algorithm that states the value between 0 and 1. It is based on the concept of odds ratio. Logistic regression uses a logistic function or sigmoid function to state its output.

XGBoost: The Boosting technique is an ensemble approach that helps to combine multiple models to fit in the best model. It helps to minimize the error by iteratively correcting. The Gradient Boosting technique increases the accuracy by training the model to predict the error of the prior applied ML model. XGBoost scores above all other algorithms in a way that the speed and accuracy increase neighbor as the parameter tuning and build in routine to impute the missing values in the dataset helps it achieve higher performance.

KNN: K-nearest neighbors is a non-parametric method used for both classification and regression. It is also known as a lazy learner algorithm. In this algorithm, the number of neighbors are selected by the user, and based on which category most neighbors belong, the new data points are classified. This algorithm is robust to noisy training data.

Random Forest: Random forest is a tree-based ensemble method used in classification models and also in regression models. The ensemble based on the set is called "base learners". Each node in the tree is constructed via recursive partitioning into two descendants based on a splitting criterion. Mean prediction of individual trees is taken as the final prediction of the response variable forming the random forest.

2.4. Imbalance learning

SMOTE learning has also been used to deal with class imbalance problems. Chawla et al. (2002) proposed the SMOTE learning technique to deal with class imbalance problems and solve them using synthetic minority over sampling techniques. Over the

last few years, many modifications and advances to SMOTE learning have been proposed. Halteh et al. (2020) proposed a SMOTE learning-based predictive model in Bankruptcy Prediction of Australian SMEs and Large Companies.

Zhang et al. (2012) proposed the Near miss under-sampling technique and Hart (1968) gave the Condensed Nearest Neighbour (CNN) method. In both these techniques, the majority of class instances that were to be kept are selected. Ivan Tomek proposed the Tomek Link method which was two modifications done to CNN. Wilson (1972) proposed the Edited nearest neighbor (ENN) technique which chose three neighbors of minority class misclassified from majority class instances and was removed. Kubat et al (1997) gave One Sided Selection which was a combination of Tomek Link and CNN as an under-sampling approach. Lastly, the Neighbourhood Cleaning rule was given by Jorma (2001) which was a combination of CNN and ENN.

2.5 Under-sampling techniques explanation

In this section, we have explained under-sampling techniques used for balancing the dataset.

Near Miss and Condensed Nearest neighbor under-sampling techniques select the majority class instances to be kept rather than removing them.

2.5.1 Near Miss under-sampling

Near Miss is an under-sampling technique that further has various variants. It uses the concept of Euclidean distance of the majority class from the minority class. Zhang et al (2012) proposed three versions of this technique which are as follows:

- Near Miss-1: In this version, the majority class instance with a minimum average distance from the closest three minority class instances is selected.
- Near Miss-2: In this version, the majority class instance with a minimum average distance from the farthest three minority classes is selected.
- Near Miss-3: In this, the majority class instance with minimum distance to each minority class instance is selected.

2.5.2 Condensed Nearest Neighbor Rule under-sampling

Condensed Nearest Neighbours (CNN) is an under-sampling technique that aims to retain the model performance without any loss. In this technique, if a point can be correctly classified by the model, it is kept else discarded. This technique was proposed by Peter Hart in 1968

In the following under-sampling techniques, majority class instances are removed instead of being selected to be retained.

2.5.3 Tomek Links for under-sampling

Tomek Links is another under-sampling technique given by Ivan Tomek as two modifications to the existing CNN method. In the first modification, searching pairs of instances from each class were suggested based on the minimum Euclidean distance amongst them. So a tomek link between two points A and B was defined if both were closest neighbors to each other and belonged to different classes. In this method, minority class instances are kept constant, while majority class instances closest to minority class are accounted as misclassified and are removed.

2.5.4 Edited Nearest Neighbors Rule for under-sampling

The Edited Nearest Neighbour under-sampling technique was proposed by Dennis Wilson in 1972. In this method, three nearest neighbors are chosen from the majority class which is misclassified as minority class and removed. In this technique, misclassified instances from minority classes are also removed. This technique generally gives better performance when combined with other under-sampling methods.

- In the following, under-sampling techniques instances of the majority class are kept and discarded simultaneously.

2.5.5 One-Sided Selection for under-sampling

The one-sided Selection under-sampling technique was proposed by Miroslav Kubat and Stan Matwin in 1997. This technique is a combination of the Tomek link and the CNN method. Tomek links remove the unclear points from the class boundary and CNN removes the redundant points which are distant from the decision boundary.

2.5.6 Neighborhood Cleaning Rule for under-sampling

The Neighbourhood Cleaning Rule (NCR) is an under-sampling technique proposed by Jorma Laurikkala in 2001. This technique is a combination of two under-sampling methods, CNN and ENN. The first method removes redundant instances whereas the second method removes unclear points. This technique first selects all the minority class instances, then removes unclear points using ENN and finally the remaining misclassified points are removed using CNN only if the size of the majority class is still larger than half of the minority class. By default, the number of neighbors is kept at three.

2.6. Challenges and limitations of the previous work

- In the previous works, the Crunchbase database has been used to predict M&A but the performance of the other models such as XG Boost, Random Forest, Logistic Regression needs to be validated and discussed in accordance with the baseline results. These predictive classification models can enhance the performance of the ML models.
- The past studies cover a few specific developed and semi-developed countries for predicting M&A, a worldwide analysis could give a larger picture of the M&A prediction in which companies from any place in the world get acquired and what factors impacted that event.
- The problem of an imbalance dataset inherited from Crunchbase data has not been handled. It has been seen that a larger chunk of the deal gets canceled due to undervaluation and other reasons so the target class has imbalance instances. This problem needs to be addressed to provide more effective models.
- Also, a combined study of the details of acquirers of companies along with the macroeconomic variables and Intellectual property of acquired companies which can yield better prediction results has yet not been done to the best of our knowledge.

2.7. Our Contributions

We have overcome the challenges mentioned in Section 2.6 by conducting a holistic study.

- We have incorporated various factors like intellectual property & Macroeconomic variables which impact the decision of the M&A event.
- We have also considered more countries than those covered previously.
- We have handled the problem of an imbalanced dataset which is inherited from the Crunchbase database and is found in real-life scenarios with the help of different under-sampling techniques.

CHAPTER 3

RESEARCH METHODOLOGY

This chapter gives details about the methodology adopted during the conduct of this project. Section 3.1 covers the description of the problem statement, Section 3.2 presents our proposed framework and the steps included to conduct it and Section 3.3 discusses how we have dealt with the imbalance dataset.

3.1 Problem statement

As we now know the importance of M&A and IPO, we can understand how important it becomes to determine whether a company should opt for M&A or IPO or is well versed in the way it is operating. There are many factors that impact the decision of selecting any one of these. In this study, we aim to predict these statuses that is whether a company should opt for M&A or they should offer IPO to raise capital for business projects or they should keep operating the way they are. For this we have deployed multiple classification models along with sampling models which balance the data to find the best performing model. Few factors taken into consideration are macroeconomic variables, corporate taxes, intellectual property, and historical data regarding funding.

3.2. Proposed Framework

The project framework for this report was subdivided into 8 steps. Below is the pictorial representation of the steps involved during the conduct of this report (See Fig 3.1). A detailed explanation of these steps is provided below and also in further sections.

Step 1: The first task was gathering the companies, acquisitions, funding rounds, investors' profiles, and degrees data from the Crunchbase data repository. Crunchbase does not provide public access to the database, so we gathered the above-mentioned datasets with updated information till 2013 from Github, and the macro-economic dataset was gathered from the World Bank Database (WB).

Step 2: After gathering the separate datasets files in csv (comma-separated value) format, next we needed to create a meta-data file that consists of unique and

meaningful attributes from these files. So a master file was created with the metadata which consisted of the attributes which would be required for the pre-processing stage and other attributes creation.

Step-3: The most important task of the ML model building process is pre-processing. The master data was used for the pre-processing stage. A detailed explanation of the steps performed during the preprocessing stage is provided in Section 2.1.

Step-4: Further label encoding and column transformer were performed in the preprocessed dataset. A detailed explanation of the steps performed during the preprocessing stage is provided in Section 2.4.

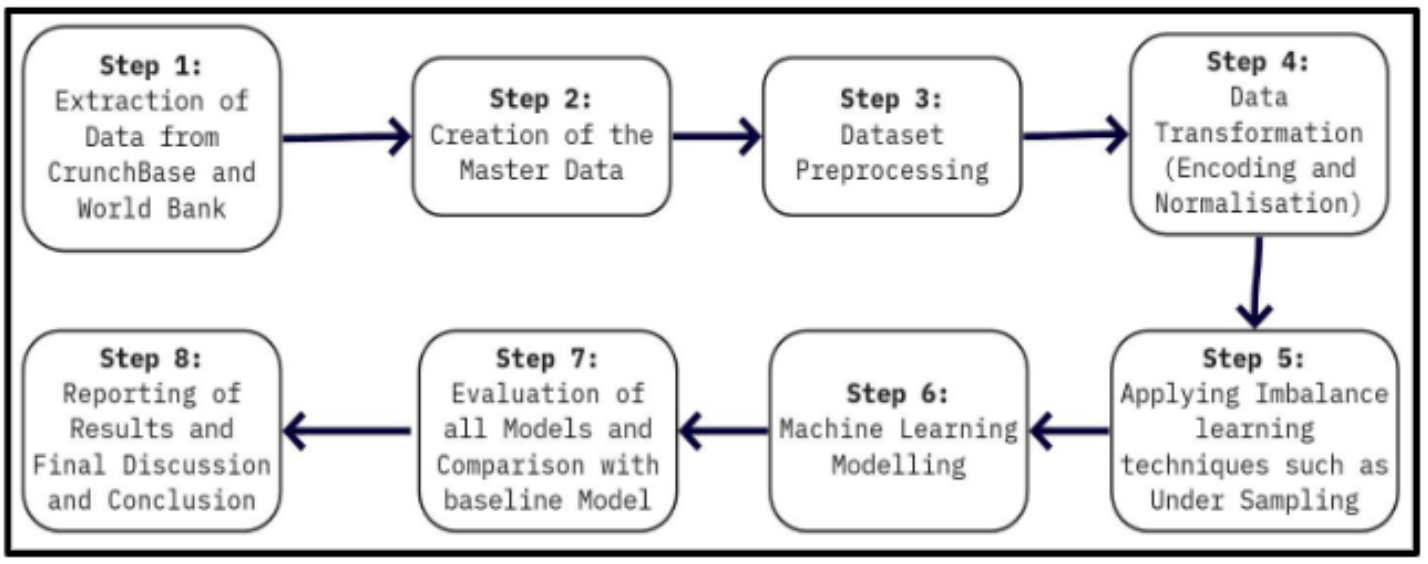

Fig 3.1: Flowchart of step involved during this report

Step-5: One of the most important steps in this research was dealing with Imbalanced classes in the target attribute. Various Techniques were applied to make data more balanced and appropriate for modeling ML algorithms in the next task. A detailed explanation of applied imbalance techniques is provided in Section 2.5.

Step-6: After applying all of the preprocessing, encoding and imbalances techniques, finally the dataset was fitting accordingly with the classification ML models.

Step-7: The Performance of the ML models was compared with the evaluation metrics described in detail in the Results section.

Step-8: Finally we have discussed the conclusion for the research conducted and discuss the future work which could be conducted in this domain area. We present the limitations of this research and suggest further work for other researchers who wish to work on this problem.

3.3 Dealing with Imbalanced Dataset

An imbalanced dataset is such where classes are not equally distributed. The class imbalance problem is prevalent in fraud detection problems where one class is more than the other class in order of 1:100.

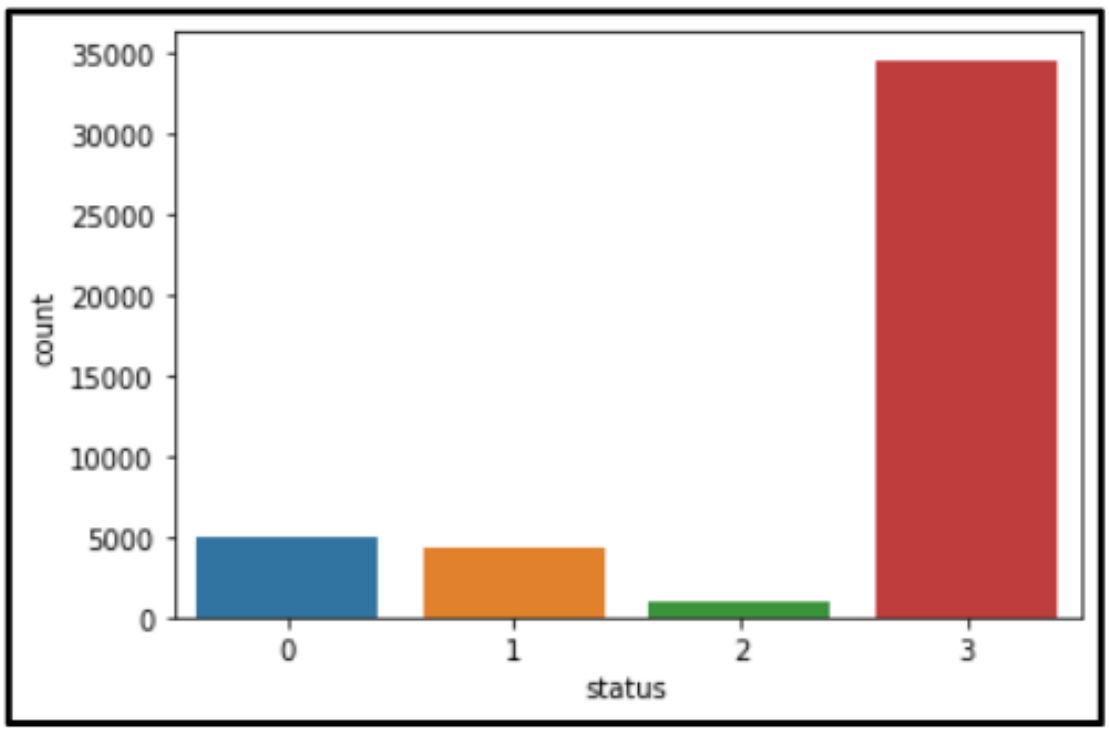

Fig 3.2: The count of the instances for the target class for the dataset used in this project

It is depicted in Fig 3.2, above that this dataset suffers from class imbalance problems. There are multiple approaches to deal with class imbalance like the resampling approach using up-sampling the minority classes using cases of majority classes or down-sampling the majority classes to match the cases in minority cases. The recent approaches like SMOTE learning, ADASYN, Random under-sampling learning are used now along with their advanced models to deal with class imbalance using the creation of synthetic newer cases and making the shape of the dataset balanced.

In this project, we have explored the different approaches for the Majority down sample techniques (under-sampling) to make the dataset balanced.

CHAPTER 4

EXPERIMENT SETUP

This section gives a detailed explanation of the dataset used, steps for data preparation, pre-processing of the data, and tools used for this project. It also explains various experiments conducted during this project.

4.1 Data Description

For this study, we have collected the data from multiple sources and then merged it. We gathered data regarding companies, details about their funding, and the number of investors from the Github repository[1] which had data extracted from the Crunchbase repository. Further, we collected data regarding macroeconomic variables and intellectual property from the World Bank data repository[2]. Finally, we collected the corporate taxes dataset for 2015. A brief description of the variables is provided in Table 4.1.

Table 4.1: Data variables/features and their description

Broad categorization of variables	Variable name	Description
About companies acquired	1. *Permalink*	Universally unique identifier (UUID) for an entity or category/location
	2. *Company name*	Names of the companies considered in the dataset
	3. *Homepage Url*	Links of the website of the companies
	4. *Country code*	Gives 3 letter abbreviation of the country name
	5. *Country Region*	Region in which company is based
	6. *Country City*	City in which company is based

[1] https://github.com/notpeter/crunchbase-data

[2] https://databank.worldbank.org/source/world-development-indicators

	7. *Primary Category*	Main vertical or sector in which company established and offering their services
About investors for different companies	8. *Funding amount USD*	Total amount of money raised from investors by a company in USD
	9. *Funding rounds*	Total number of funding rounds conducted by a company
	10. *Founded year*	Year in which the company was founded
	11. *First funding at*	Date on which first funding for a company took place
	12. *Last funding at*	Date on which last funding for a company took place
	13. *Funding Duration*	Duration for which funding was conducted for any company. It is the difference between the last funding and first funding date.
	14. *Number of investors*	Number of investors for each company
Macroeconomic variables	15. *Consumer price index (2010 = 100)*	The consumer price index for each given country in 2015. It tells us about the price changes linked with the cost of living.
	16. *Foreign direct investment, net (BoP, current US$)*	Foreign direct investments for each country to which company belongs in 2015. It tells us the investment in a country by other firms or individuals.
	17. *GDP growth (annual %)*	Gross domestic product growth in 2015 for the country to which the company belongs.
	18. *Inflation, consumer prices (annual %)*	Inflation helps us assess price changes in a country for the year 2015 to which the company belonged.

	19. *Real interest rate (%)*	Real interest rate helps in determining the purchasing power of the value of interest on any investment for the year 2015 for the country in which the company is based.
	20. *Unemployment, total (% of the total labor force) (national estimate)*	This variable tells us about the people that are currently unemployed but looking forward to a job in a country in 2015.
Taxes	21. *Corporate taxes 2015*	Corporate taxes in different countries in which the company is located for the year 2015
Intellectual Property related variables	22. *Patent applications, residents*	Total number of patent applications filed in the country in which company is located the in year 2015
	23. *Patent applications, nonresidents*	Total number of patent applications filed in the country by people outside that country till 2015
	24. *Trademark applications, total*	Total trademark applications in a country till 2015
Dependent Variable	25. *Status*	This tells us the status of the company which can be any one of four: acquired, closed, ipo or operating.
Total	*26*	In our dataset, there are 44682 rows and 26 columns.

4.2 Data pre-processing

This section covers a brief explanation of the preprocessing steps we have done to prepare our dataset for model building and selection.

4.2.1 Data cleaning

The dataset consisted of hashes, hyphens, and dashes and it was essential to replace such entries with blank spaces and then apply the techniques of preprocessing because sophisticated techniques of machine learning could not be applied on such a dataset.

4.2.2 Removing duplicate entries

The dataset consisted of some of the duplicate rows and the same records are not essential for modeling purposes. The model can learn from a single unique record, so those duplicate data were removed from the further versions of the dataset.

4.2.3 Correcting structural errors

The dataset consisted of the same country names and regions written in different ways like Saint Louis was written at St. Louis at some places or Bangalore was written as Bangalore city. To make the data uniform such structural errors were removed.

4.2.4 Missing values

The dataset with missing values does not add any value to the models deployed on it. It's necessary to treat the attributes with such values using appropriate ways. The researchers use various approaches to deal with missing values like:

- Deleting the rows which contain the missing records, or
- Replacing the missing values with the mean values or with the median or mode value.
- The fill method is sometimes used to create a new class or category for the missing records.
- The regression model is fitted onto the data and then the values are predicted for the same.

The M&A dataset used in this project consisted of missing values and to impute values into them we applied many techniques corresponding to the nature of the attribute.

- The attributes which didn't add any meaning to the dataset preprocessing and modeling purpose were removed from the dataset. Such attributes are: *permalink of the company, name of the firm, homepage_url of the companies, firm's working state_code, region in which firm is located, city in which firm is located, founded date of the company, first_funding date of the company, last_funding date of the company, secondary category in which form is operating, tertiary category in which form is operating, other category in which form is*

operating, first_funding date in UTC format, last_funding date in UTC format, year in which last funding was received.

- The missing primary *category of the firm* was replaced with the 'other' class category.
- The missing values in the *country* attribute were replaced with the 'other' class category.
- The missing values in the *total funding received by a firm* were imputed with the mean values of the funding amount of the category of those firms.
- The missing values in the other *macro-economic* attributes in the dataset were replaced with the mean values of the non-null records.
- Finally, the dataset had the attributes which added meaning to the modeling purpose and also now none of them had any missing values.

4.2.5 Encoding of Data

To encode the categorical values various approaches are adopted like the 'find and replace' approach where we explicitly assign and encode the categorical values as an integer value. The other approach is Label Encoding, where each value is converted into a number. The other approach is One-Hot-Encoding, which creates dummy variables with the values of 0 and 1 for each of the classes of the primary attribute.

Other approaches can also be tried, but for the categorical attributes like *current status of the company (as of 2015), work country of the company,* and *the primary category of the company* had categorical (character) values in them.

The preprocessing library offers the LabelEncode method to transform these 3 attributes and create numerical values instead of categorical classes.

4.2.6 Normalization of Data

The dataset used in this project was normalized for better model building using Min-Max Scaler. We have applied the Min-Max scaler to normalize the features into the 0-1 range.

$$X\,\text{scaled} = \frac{X - \min(X)}{\max(X) - \min(X)} \qquad (1)$$

where x represents the single feature vector.

Thus the dataset now had all of the normalized attributes with values ranging 0-1, which would be beneficial for modeling purposes.

4.3 Evaluation Metrics

Here, we have explained evaluation metrics used to compare performance of our machine learning models.

I. **Micro F1:** In this case, F1 is calculated using total true positives, false positives, and false negatives of all classes combined.

II. **Macro F1:** In this metric, F1 scores will be calculated for each class and then their average will be calculated. In this, each class is given equal importance.

III. **1- Hamming loss:** Hamming loss is the hamming distance between predicted and actual values. The lesser the distance, the better is our model. Its values range from 0 to 1. Hamming loss discusses the error or the incorrect prediction, so 1- Hamming loss tells us about correct predictions and the accurate results.

4.4 Tools used

- MS Excel
  - For merging of the Crunchbase Company and other datasets with the macro-economic data from IMF, Ms. Excel PowerQuery was used in this project.
  - Also for Data Visualization and Exploratory data analysis, the Excel tools were used.

- Jupyter notebook and Spyder(Python)
  - After the creation of the master dataset, the preprocessing and ML fitting models were deployed in Python (Jupyter notebook and Spyder).

- The Data Visualization and Exploratory data analysis were also done through the python libraries like matplotlib, ploty, seaborne, and others.

- Power BI
  - Power BI stands for Power Business Intelligence Toolkit.
  - For Data Visualization and Exploratory data analysis,modeling the Power BI desktop version was used.
- Tableau
  - For Data Visualization and Exploratory data analysis the Tableau desktop version was used.

4.5. Experimental Results

We have performed exploratory data analysis with respect to the attributes in our dataset and then we shared data modeling results.

4.5.1 Exploratory data analysis

This section covers the different insights we have drawn from the dataset using the tools we have discussed in Section 4.3.
Exploratory data analysis is a method of summarizing our dataset and getting useful insights between the variables with the help of data visualization and hypothesis testing. It helps us understand the data beyond the modelling part. In this project, we have conducted the exploratory data analysis using the following techniques.

Correlation matrix: To find the type of relationship between variables, we have used a correlation matrix. This matrix returns correlation coefficients between two variables and helps us understand the behavior of one variable concerning the other.

Bar plots: It is a data visualization technique which is a chart or a graph. It represents aggregates or summaries of the categorical data based on numeric values. In this chart, the data is presented using rectangular bars which can be placed horizontally or vertically.

Bubble charts: This is another data visualization technique that is used to represent data in two to four dimensions. The first two dimensions are treated as coordinated on the axis, the third dimension is the color and the fourth dimension is represented by the size of bubbles plotted.

TreeMap: It is a data visualization technique used for hierarchical data which is represented in the form of rectangles. It is relatively easy to interpret and draw insights from them.

Now we would be presenting our EDA insights using the tools discussed earlier.

I. Number of companies belonging to each status type

Fig 4.1 and 4.2 provide us with an analysis about the number of instances for each of the status types in our dataset. The Bar Plot and the Map representation give us a better idea to know about our target variable.

(a) Barplot

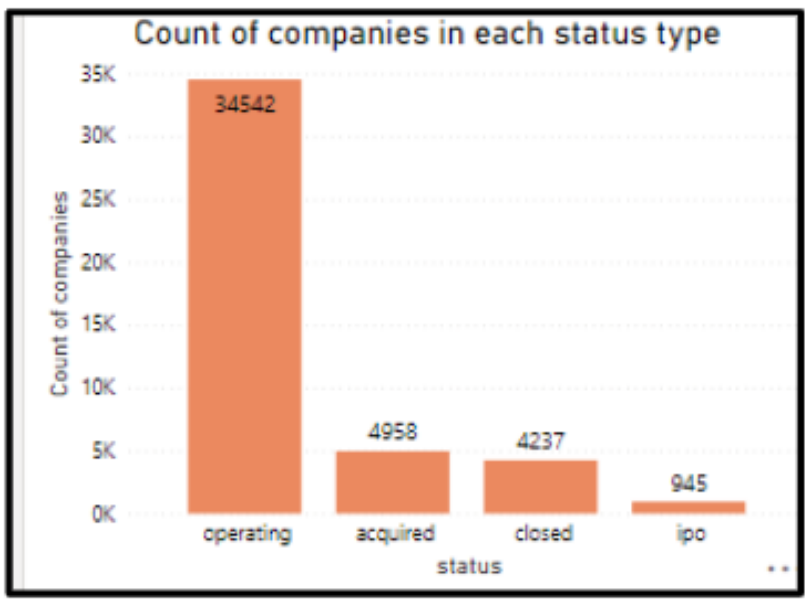

Fig 4.1: Number of companies in each status type

Interpretation: As we can see fro the Fig 4.1, the dataset consists of total of 44682 companies out of which 34542 companies are operating as of 2015, 4958 companies have been acquired, 4237 companies have been closed and 945 companies offered initial public offerings to raise money for their business as of 2015.

(b) Map representation of a status of companies in different countries

Fig 4.2: Map representation of the status of companies

Interpretation: Fig 4.2 shows the countries in which the companies mentioned in the dataset belong along with the status of companies mostly found. As we can see many companies are mostly located in South America, some parts of Asia and Africa. Also, companies in Asia have more diverse status as compared to other places.

II. Relation between various variables using correlation

Fig 4.3 shows the correlation between different variables in the dataset. According to the correlation matrix, we can observe the following:

- Funding rounds are positively moderately correlated to the number of investors and funding duration for the companies. The values are 0.63 and 0.75 respectively.
- There is a high negative correlation between foreign direct investments and non-resident patent applications. The value is -0.91.
- Patent applications of non-residents are moderately positively correlated to corporate taxes which is 0.78.

- Unemployment percentage and foreign direct investments are positively correlated but the correlation is low, that is 0.37.

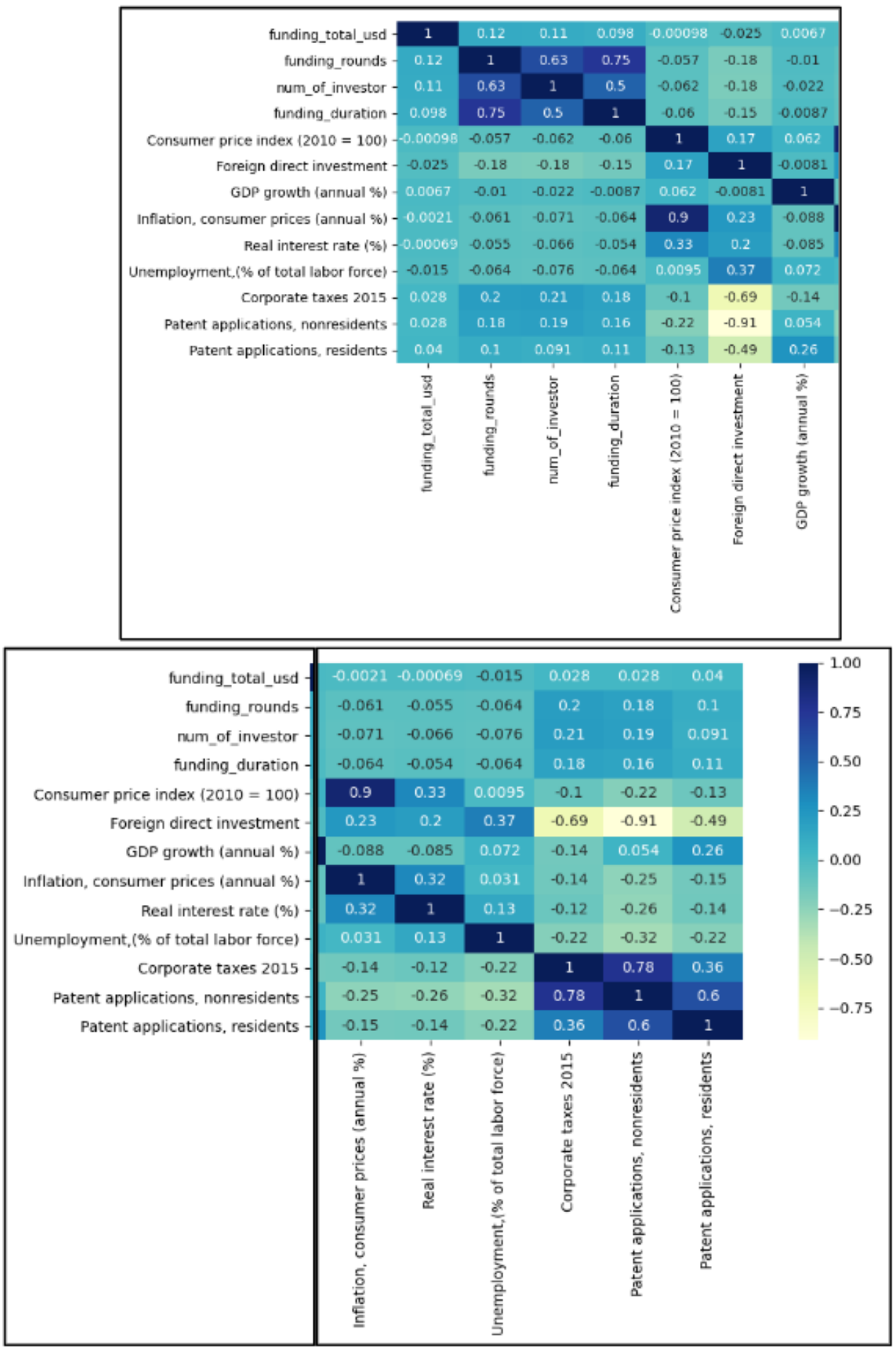

Fig 4.3: Correlation between macroeconomic variables and variables related to the acquisition

III. Top 10 Sector across various categories

Interpretation: As per Fig 4.4, the top 10 sectors in our dataset are Software with 2759 companies, biotechnology with 2664 companies, followed by others which are companies from unknown sectors 2419.

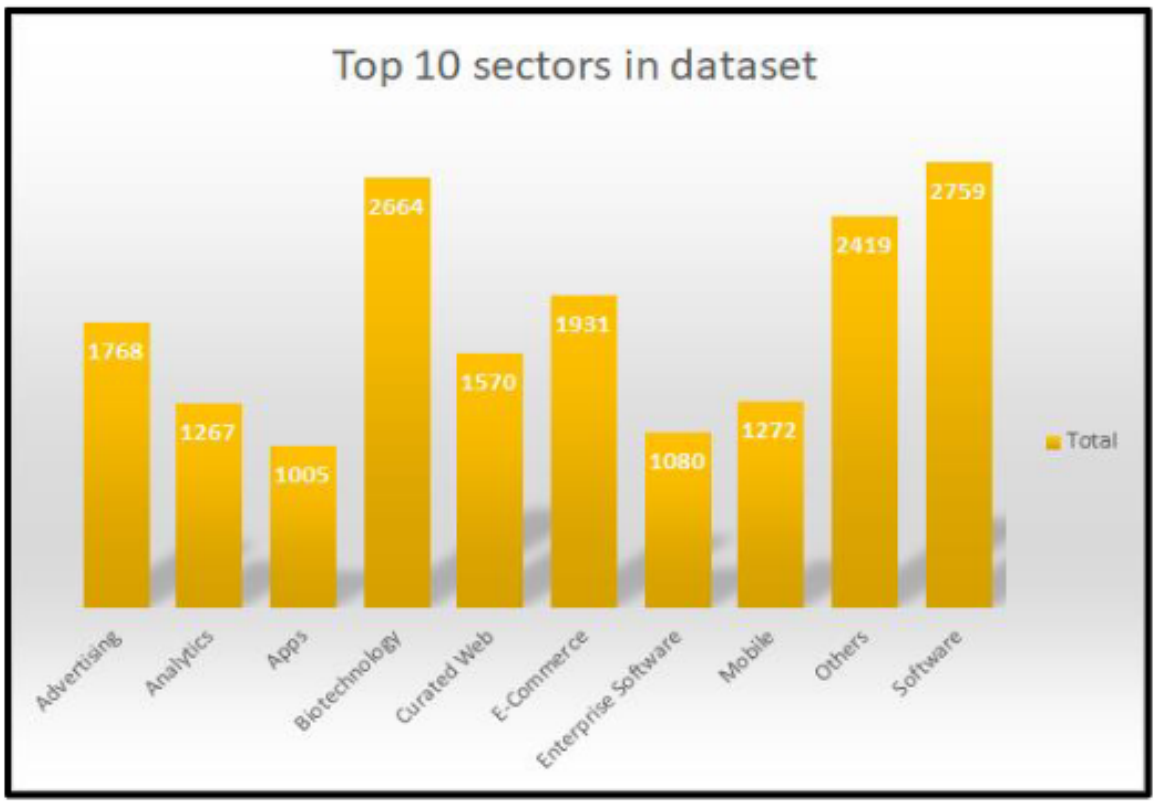

Fig 4.4: Top 10 sectors in our Dataset

E-commerce with 1931 companies, Advertising firms with 1768 instances, 1570 Curated web firms, 1272 mobile companies, 1267 Analytics companies, 1080 Enterprise firms, and finally firms related to Apps with 1005 instances.

IV. Top 10 sectors in every status type

Interpretation: According to Fig 4.5, Software companies followed by advertising companies were most acquired.

Similarly, Companies that were closed belonged to curated web categories or categories other than those given in the dataset. Companies belonging to the biotechnology category offered the most IPO followed by the software and Health care sector. Lastly, the companies which were still operating belonged the most to the software, biotechnology, and e-commerce category respectively.

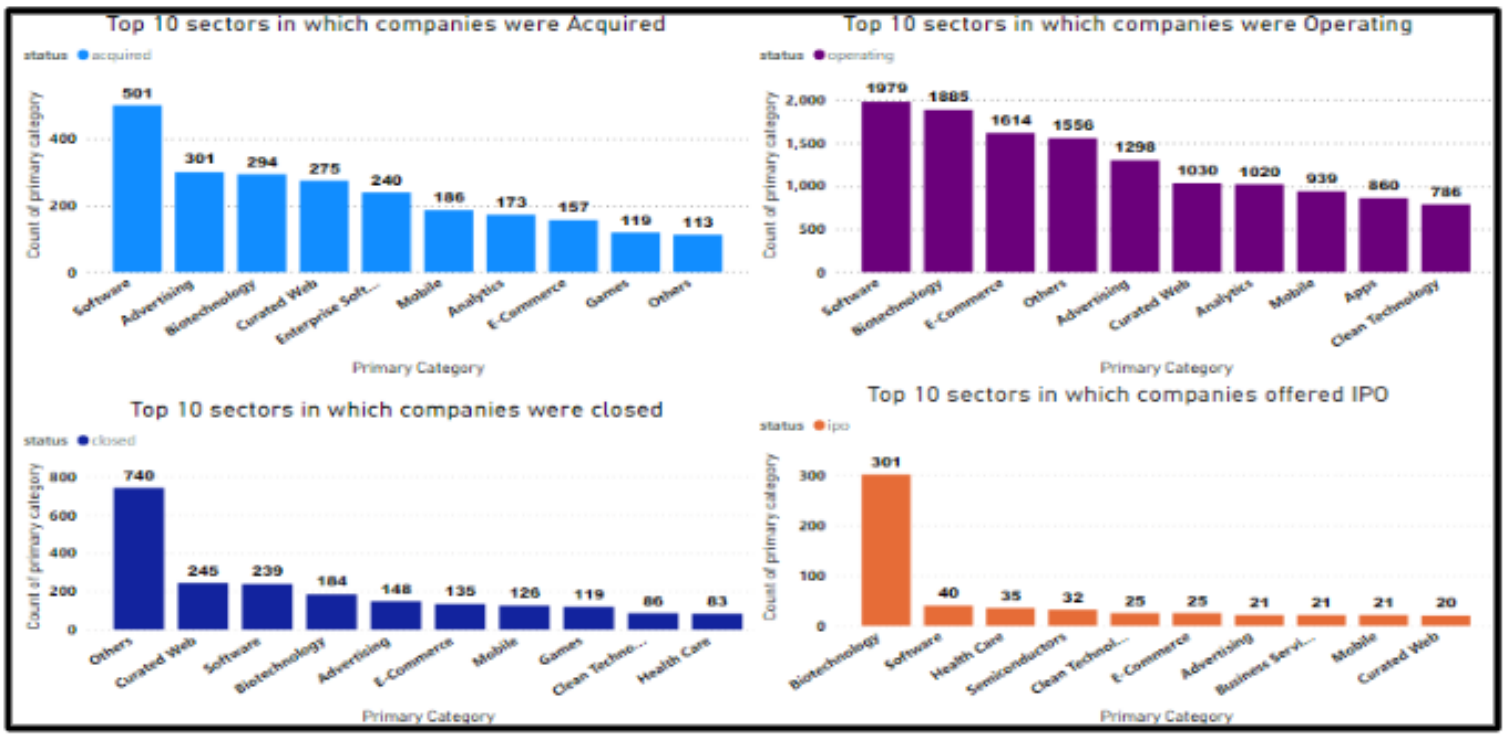

Fig 4.5: Top 10 sectors in every status type

V. Top 5 countries in Various Categories

Interpretation: From Fig 4.6, we can observe that the USA is the first and foremost country that has companies with the most acquisition, firms that offered IPO, firms that closed, and lastly the firms which were operating. The reason behind this is that our dataset consists of 44682 companies out of which more than 23000 companies are based in the USA. The second-largest acquirer country is unknown, followed by Canada.

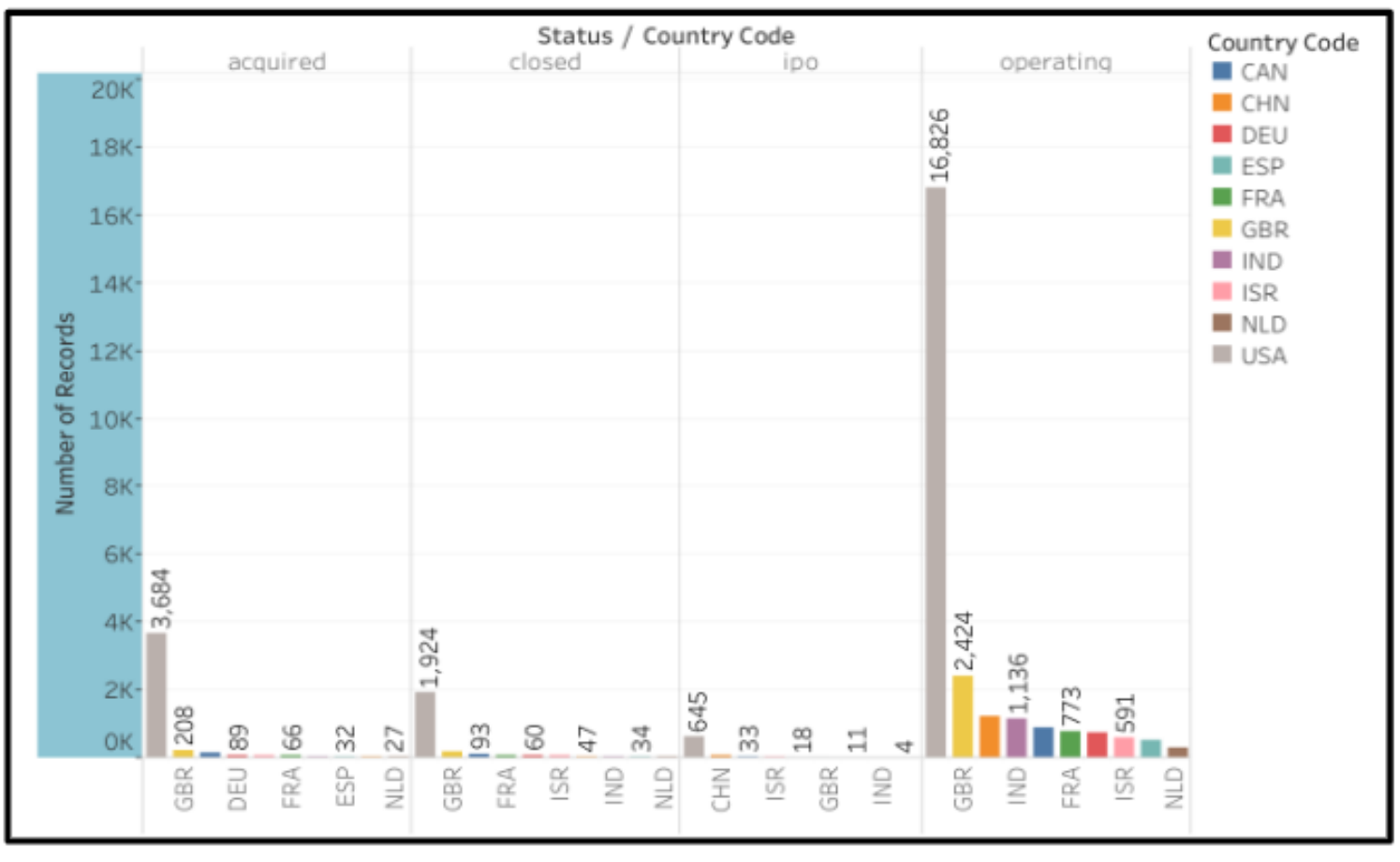

Fig 4.6: Top 5 countries in Various Categories

Similarly, the second largest country to offer an IPO is China with 73 entries. There are 1157 entries for unknown countries where companies were closely followed by the United Kingdom of Great Britain with 200 instances. Lastly, unknown countries have the second-most instances where companies were operating followed by Great Britain and China respectively.

VI. Highest funding in top few sectors

Interpretation: According to Fig 4.7, we can see that out of all the sectors given, the communication sector has received the the highest funding of cumulative $30.1bn followed by companies operating in field of semiconductors with the funding of $17.6bn and companies in the automotive sector with funding of $8.2bn. This figure shows us the top sectors with high cumulative funding amounts.

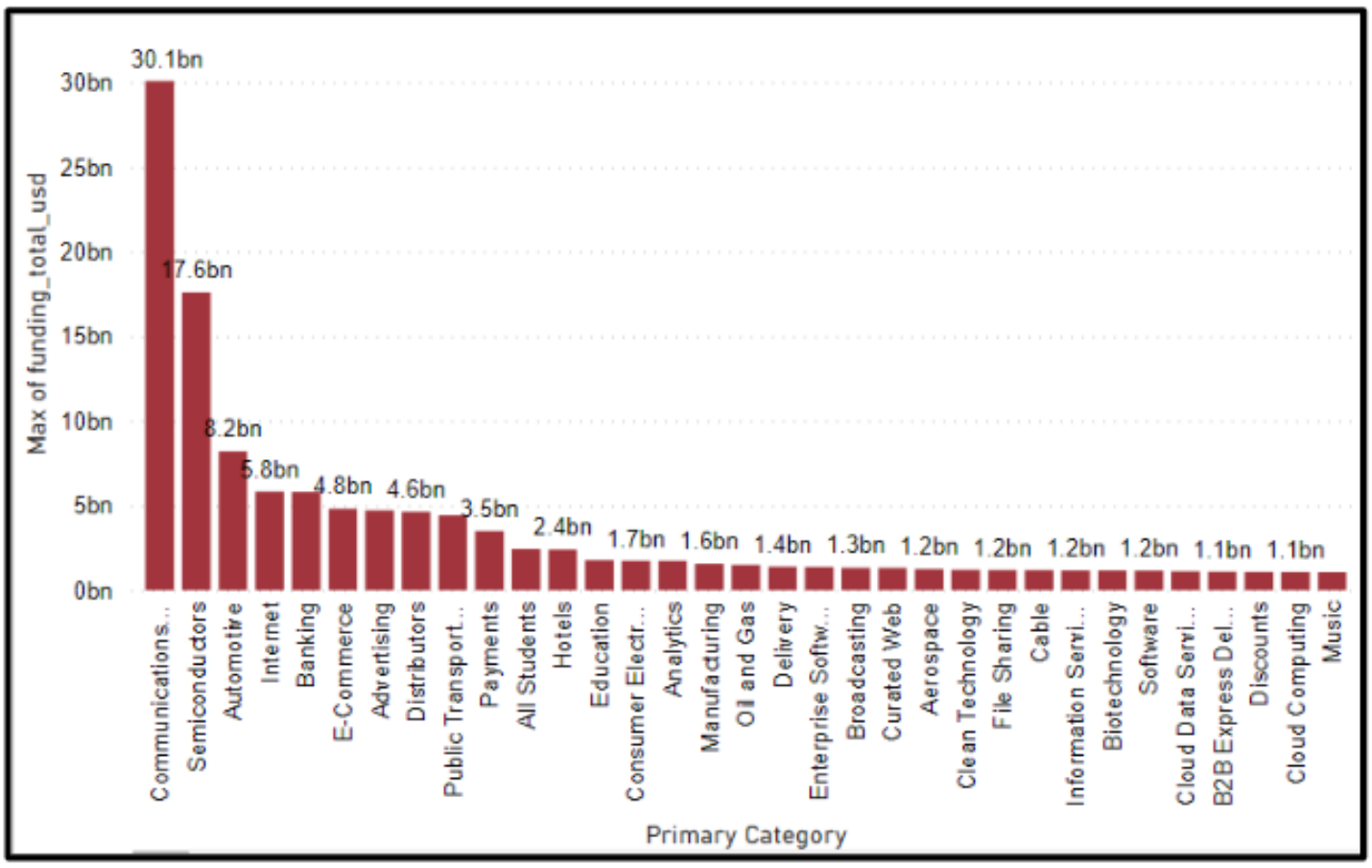

Fig 4.7: Highest funding in top sectors

VII. Count of Records of the Acquired Companies Categories

Interpretation: As depicted by the Fig. 4.8 Bubble chart for the count of acquired companies categories, we can analyze that the Software industry with a count of 1465 has been acquiring the most number of the companies, followed by Advertising with 1011 and then others such as Biotechnology, Curated Web, Enterprise Software, E-commerce and many more.

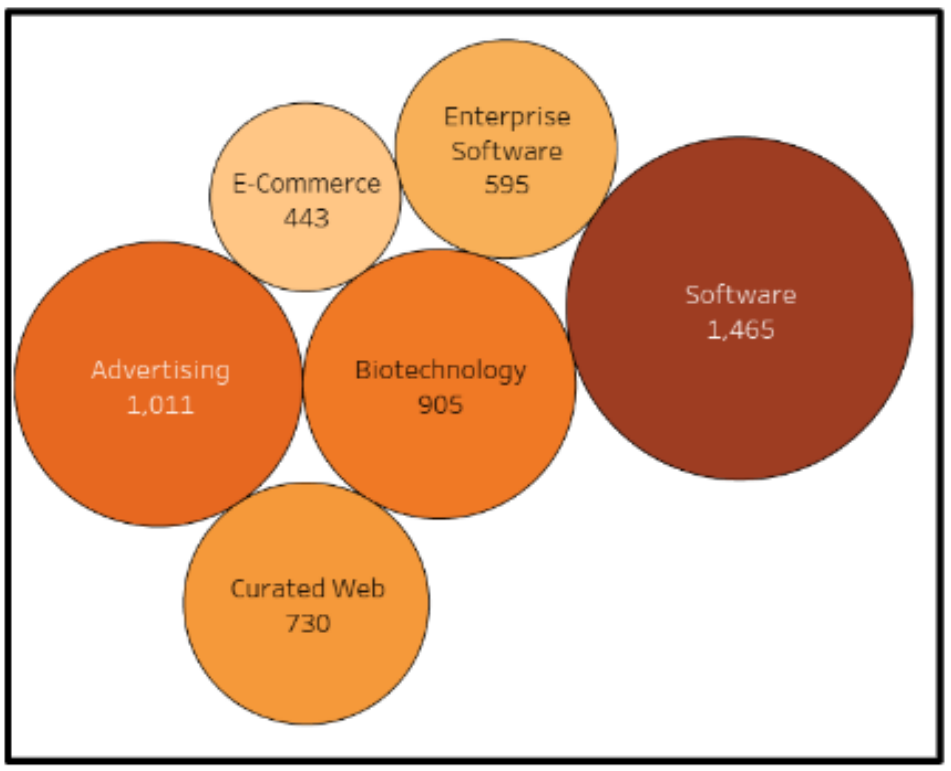

Fig 4.8: Bubble Chart representation of the Count of Records of the Acquired Companies Categories

VIII. Records of the Acquired Companies Names

Fig 4.9: Bubble Chart representation of the Records of the Acquired Companies Names

Interpretation: As depicted by the above Bubble chart in Fig 4.9, for the count of acquired companies names, we can analyze that Cisco with a count of 193 have been

acquiring the most number of the companies, followed by Google with 176 and then others such as Microsoft, IBM,Yahoo! and many more.

IX. Acquisition number between the acquirer and the acquirer sector

Interpretation: As depicted by the Fig 4.10 for the count of acquired companies' categories acquiring which categories companies. The Biotechnology industry has been acquiring the other biotechnology firms most followed by Software acquiring software firms and advertising acquiring the same other advertising firms.

Acquirer_Primar..	Company_Prima..	
Advertising	Advertising	256
	Curated Web	94
Biotechnology	Biotechnology	521
	Health Care	71
Clean Technology	Clean Technology	65
Curated Web	Curated Web	127
E-Commerce	E-Commerce	89
Enterprise Software	Enterprise Soft..	69
	Software	118
Finance	Finance	56
Games	Games	76
Mobile	Mobile	63
Semiconductors	Semiconductors	106
Software	Enterprise Soft..	87
	Software	388

Fig 4.10: Count of the acquirer category targets

Similarly, the same pattern is visible for other industry companies too. But some Curated Web development companies have also been acquired by advertising firms, similarly, Enterprise software companies being acquired by advertising firms, and lastly Biotechnology firms acquiring the other Health Care firms.

X. Acquisition deal amount between the acquirer and the acquirer

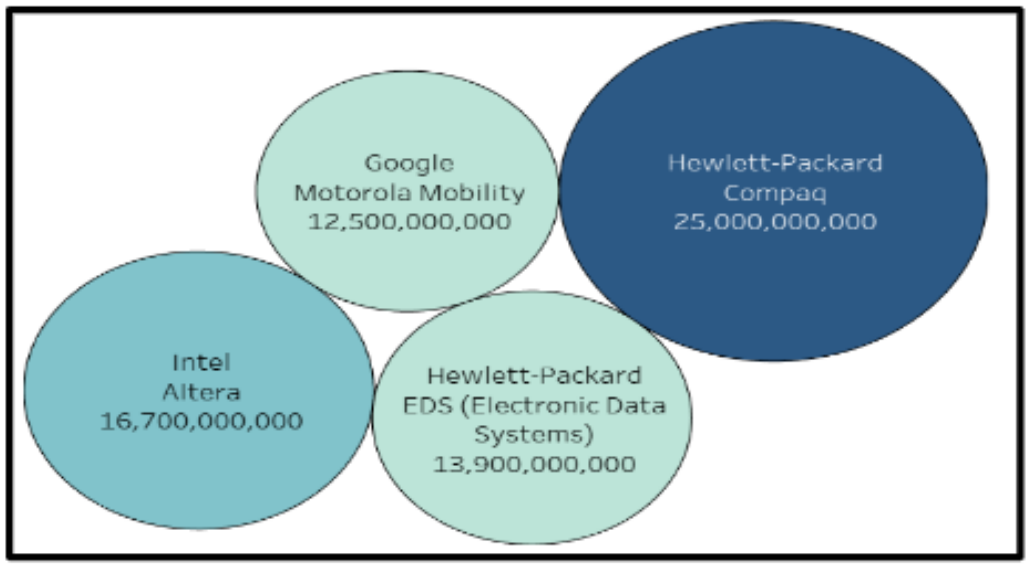

Fig 4.11(a): Bubble Chart represent the acquisition deal amount between the acquirer and the acquirer

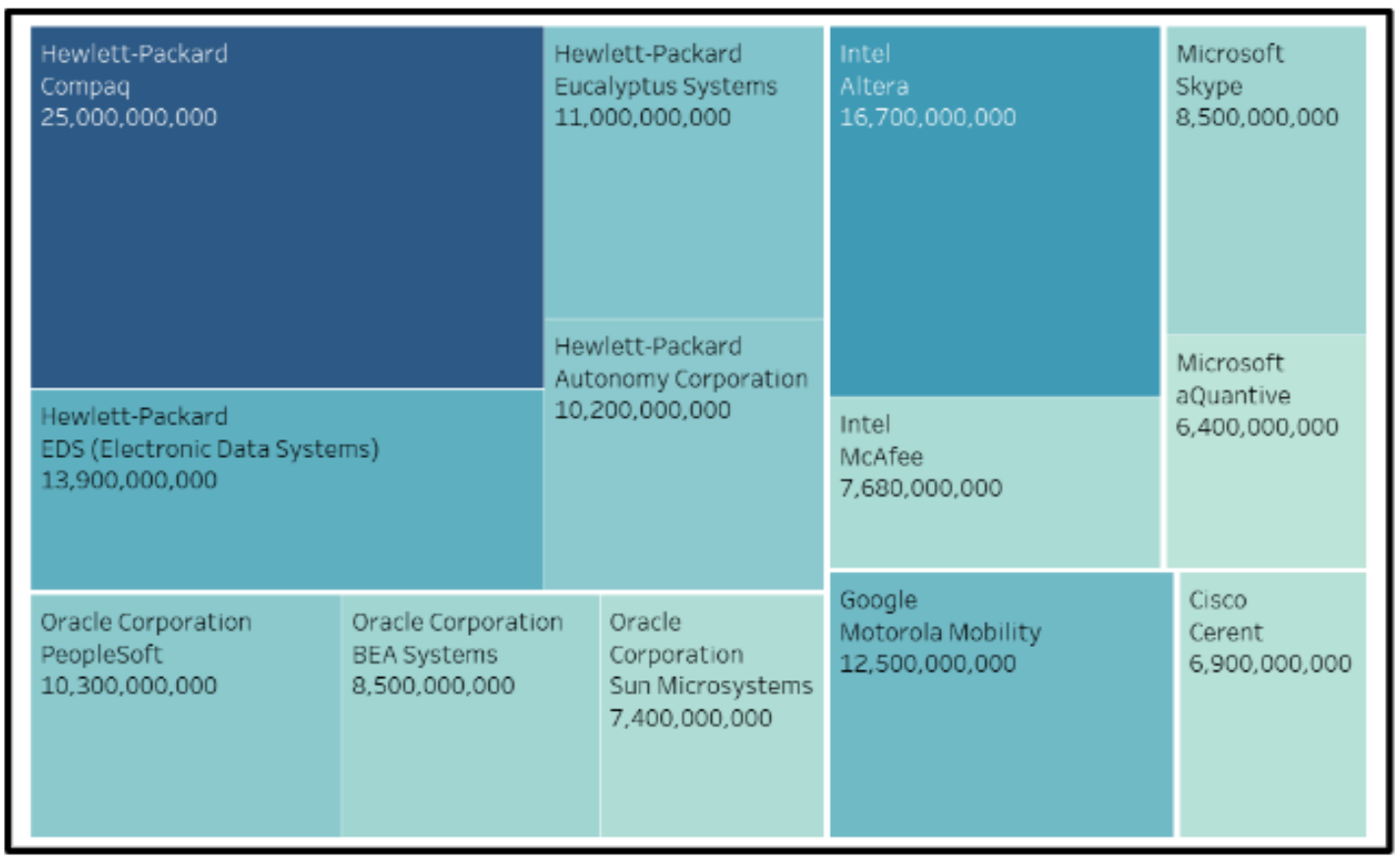

Fig 4.11 (b) TreeMap Chart represent the acquisition deal amount between the acquirer and the acquirer

Interpretation: As depicted by the above TreeMap and Bubble chart in Fig 4.11, we can see that HP acquisition of Compaq for $25 billion was the highest acquisition deal for our dataset, followed by Intel and Altera deal of $16.7 billion and HP and EDS deal of 13.9 billion and Google and Motorola Mobility for $12.5 billion. The other prominent and notable deals are mentioned in the TreeMap chart.

4.5.2 Prediction of M&A using ML Techniques

The models are fitted on the whole of the dataset using the Repeated Stratified K fold cross-validation procedure, and the value of K was taken as 10. Further, the results were obtained for the baseline model (or the dataset without addressing the issue of imbalance learning) and then for the models in which the target class was balanced using the methods discussed in Section 2.5.
The following are the results obtained for the approaches adopted in this project.

- **Evaluation Methodology: Cross-Validation Method**

For the final performance evaluation the dataset is split into K (10) sets or folds and the K-1 (9) folds are used as training data and the remaining 1 set is used for testing purposes (see Fig 4.12). Similarly, this process is repeated over K times and every K set is used for training and testing purposes and finally, the mean score is taken as performance evaluation score or Cross Validation score.

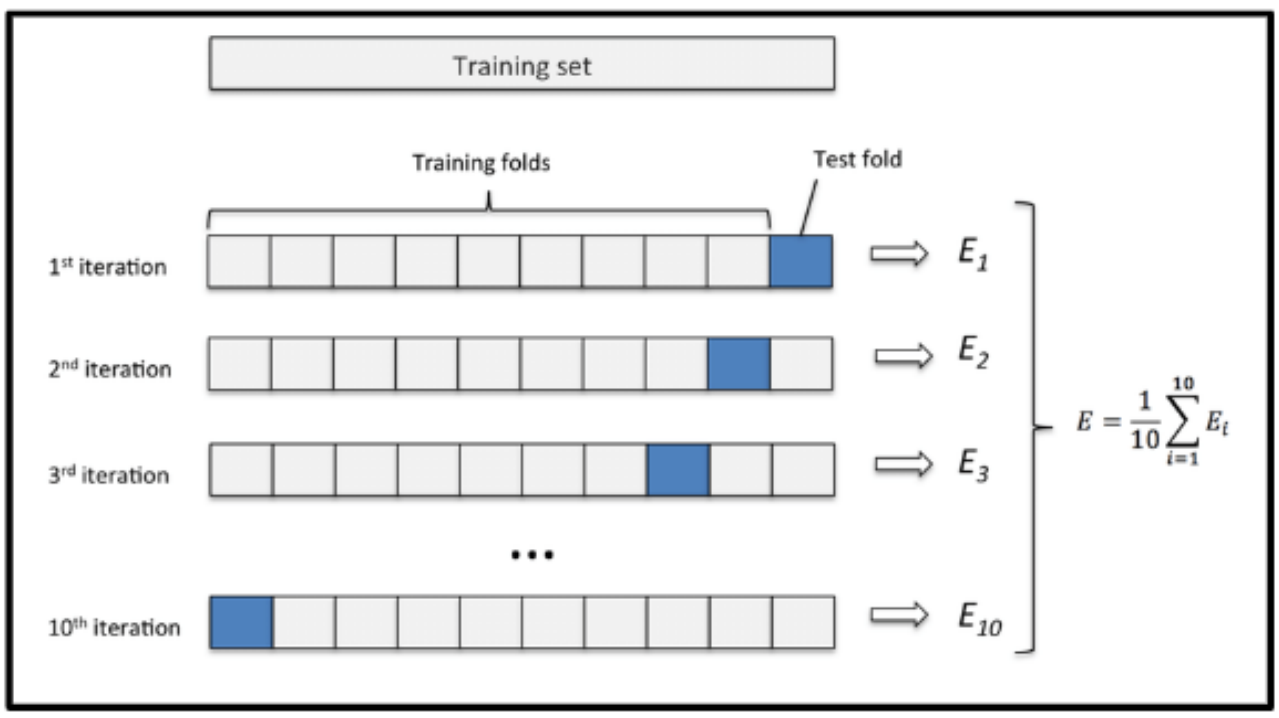

Fig 4.12: K(10) fold cross-validation Diagram

(Source: Rosaen. (2016))

An alternative method of splitting the dataset into training data and test data is mostly followed by the researchers. The training dataset contains a major chunk of data with the known output and the model is fitted and learns the pattern on this data to predict the unknown instances. The test dataset is the remaining data that is used to test our

model's prediction on this subset and validate the actual labels. The different split ratios of train-test split like 60:40, 70:30, 80:20, and 90:10 are taken by researchers to obtain the results.

In this project, we have split our dataset based on repeated K-fold cross-validation to evaluate the performance of the models. This technique improves the estimated performance and fitting accuracy of a machine learning model. The cross-validation procedures are repeated multiple times and further the mean result across all folds from all runs is reported for final discussions. The calculated mean result is a more accurate estimate of the test or the unknown and new data points. It finally increases the mean performance and efficiency of the model on the dataset which can be also validated by the calculation of the standard error.

Here, experiments are performed on CrunchBase and WorldBank datasets without sampling and this is chosen as our baseline model.

4.5.2.1 Comparison of various ML approaches with the baseline using undersampling techniques

In Tables 2-4, RUS denotes Random under-sampling approach, OSS denotes One-Sided Selection approach, ENN denotes Edited Nearest Neighbour approach and NCR denotes Neighbourhood clearing rule approach.

Table 4.2. Results obtained for micro- F1 metric

Sampling techniques	Logistic Regression	KNN	Random Forest	XGBoost
Without sampling	0.773	0.7063	0.7479	0.7728
RUS	0.5492	0.529	0.5673	0.5985
OSS	0.832	0.8131	0.8259	0.8372
ENN	0.9425	0.956	0.954	0.9545
Tomek Link	0.822	0.7806	0.8107	0.8237
NCR	0.9228	0.9432	0.9377	0.9326

Interpretations: The results shown in Table 4.2 for the micro F1 score show that in the baseline approach i.e, without sampling, Logistic Regression and XGB were able to give the best results with a 0.773 micro F1 scores. But as discussed earlier in Section 2.4, we applied the under-sampling techniques to obtain better and more accurate results. For RUS, OSS and Tomek Link under-sampling techniques, XGBoost with micro F1 score 0.5985, 0.8372, and 0.8237 respectively was able to outperform other algorithms. For ENN and NCR under-sampling techniques, KNN with micro F1 score 0.956 and 0.9432 respectively was able to outperform other algorithms. Finally, the best under-sampling was ENN and the best algorithm was XGB and KNN with the best results among others.

Table 4.3. Results obtained for macro- F1 metric

Sampling techniques	Logistic Regression	KNN	Random Forest	XGBoost
Without sampling	0.2179	0.319	0.308	0.3013
RUS	0.3101	0.3781	0.4282	0.4523
OSS	0.227	0.3847	0.3434	0.3365
ENN	0.2451	0.709	0.6115	0.6438
Tomek Link	0.2255	0.3566	0.3366	0.326
NCR	0.2399	0.6299	0.567	0.5166

Interpretations: The results shown in Table 4.3 for the macro F1 score show that in the baseline approach i.e, without sampling, KNN was able to give the best results with a 0.319 macro F1 scores. But as discussed earlier in Section 2.4, we applied the under-sampling techniques to obtain better and more accurate results. For RUS, XGboost with a macro F1 score of 0.4282 was able to outperform other algorithms. For OSS, Tomek Link, ENN, and NCR under-sampling techniques, KNN with macro F1 score 0.3847, 0.3566, 0.709, and 0.6299 respectively were able to outperform other

algorithms. Finally, the best under-sampling was ENN and the best algorithm was XGB and KNN with the best results among others.

Table 4.4. Results obtained for 1-hamming loss metric

Sampling techniques	Logistic Regression	KNN	Random Forest	XGBoost
Without sampling	0.773	0.7048	0.747	0.772
RUS	0.5492	0.5286	0.5624	0.5924
OSS	0.839	0.817	0.839	0.845
ENN	0.942	0.956	0.943	0.955
Tomek Link	0.822	0.782	0.822	0.823
NCR	0.923	0.944	0.923	0.941

Interpretations: The results shown in Table 4.4 for the 1-hamming loss score shows that in the baseline approach i.e, without sampling, Logistic Regression and XGB were able to give the best results with 0.773 1-hamming loss. But as discussed earlier in Section 2.4, we applied the under-sampling techniques to obtain better and more accurate results. For RUS, OSS and Tomek Link under-sampling techniques, XGBoost with 1-hamming loss of 0.5924, 0.845, and 0.823 respectively were able to outperform other algorithms. For ENN and NCR under-sampling techniques, KNN with 1-hamming loss of 0.956 and 0.944 respectively were able to outperform other algorithms. Finally, the best under-sampling was ENN and the best algorithm was XGB and KNN with best results among others.

From Table 4.2, 4.3, and 4.4, we can observe that the Edited Nearest Neighbour under-sampling approach (ENN) followed by Neighbourhood Cleaning rule (NCR) were able to perform better than other approaches and the outperform ML model of XGB and KNN were able to outform other fitted models and obtain the better results.

The best sampling technique ENN along with KNN model was used for further analysis to study the impact of incorporating the additional variables like macroeconomic dataset of the concerned countries along with the intellectual property variables. The analysis is discussed in the next Section 4.5.2.3.

4.5.2.2 Variation of the value of k

In this section, we have presented an analysis on the optimal value of k for the KNN model which has shown the best results as discussed in Section 4.5.2.1.

Table 4.5. Results obtained for different values of 'k' in KNN model

ENN - KNN	k = 3	k = 5	k = 10	k = 20
micro-f1	0.95629	0.952549	0.948788	0.946506
macro-f1	0.719984	0.639429	0.478929	0.410991
1-hamming loss	0.956	0.950815	0.948708	0.9466

Interpretations: The results obtained for different values of the nearest neighbour (k) are presented in Table 4.5. The 3 nearest neighbours show the best results and the other values of the 'k' i.e. 5, 10, 20 do not show encouraging results. So the KNN model was applied with the value of the 'k' being taken as 3, throughout the conduct of the project.

4.5.2.3 Impact of incorporating macroeconomic variables and intellectual property

In this section, we have shown the performance of our best model obtained earlier on different sets of data. Here, we have selected four types of datasets:

1. Crunchbase data
2. CrunchBase data along with Macroeconomic variables
3. CrunchBase data with Intellectual property variables

4. CrunchBase data with macroeconomic and intellectual property variables

The performance is measured using the evaluation metrics values of micro - F1, macro - F1 and 1 - hamming loss.

Table 4.6. Results obtained evaluation metrics on different datasets

Datasets	micro-f1	macro-f1	1-hamming loss
CrunchBase data	0.956111	0.718068	0.954389
CrunchBase and macroeconomic variables	0.956287	0.716646	0.95432
CrunchBase and Intellectual property variables	0.955813	0.708308	0.954412
CrunchBase and macroeconomic variables + Intellectual Property variables	0.95629	0.719984	0.956

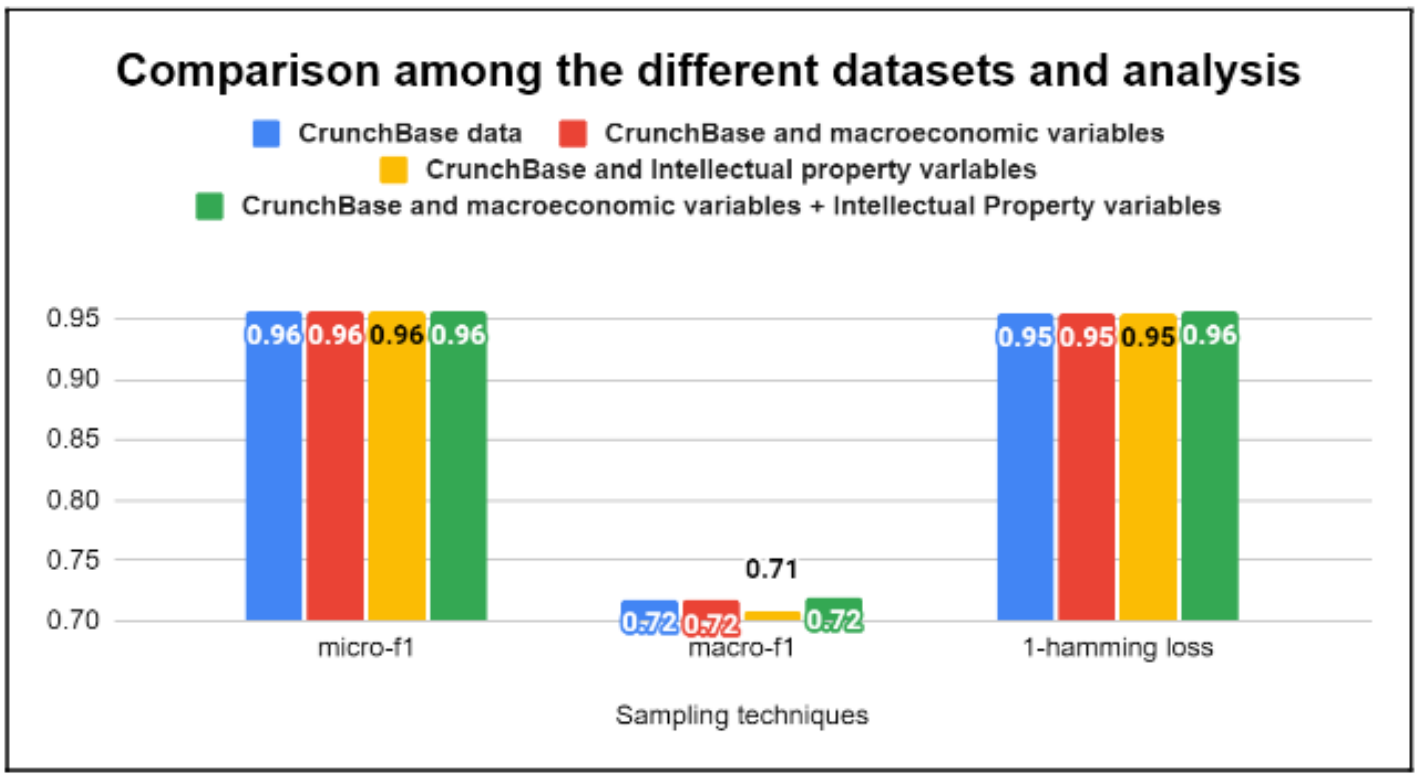

Fig 4.13: Chart depicting the comparison among the different dataset on the evaluation metrics

Interpretation: The ENN sampling technique was used for dealing with the class imbalance problem for the target class. As discussed in the Section 4.5.2.1 the KNN model outperformed the other classification models so, The ENN-KNN model was deployed on the various datasets. The results obtained in the Table 4.6 and also the Fig 4.13 shows the CrunchBase dataset along with the macroeconomic variables and Intellectual Property variables. This dataset shows the best performance in the terms of the micro-F1, macro-F1 and 1-hamming loss scores of 0.95629, 0.719984 and 0.956 respectively.

CHAPTER 5

FINDINGS AND RECOMMENDATIONS

EDA: The Exploratory Data Analysis for the chosen M&A dataset met with the defined objectives for the conducted project. Tools such as Ms. Excel, Tableau, Power BI, and python presented useful insights of the data visualization of the dataset. The maximum number of records of the dataset is from the United States and it can be validated as all the major big shot companies have headquarters in the US. Further, it was observed that the major acquiring firms are operating in the Biotechnology and Advertising and Web-domain categories. The top firms like Google, IBM, Cisco have been involved in the biggest M&A events in the past decade. Further communication sector followed by semiconductor sector received the highest funding, Cisco acquired the most companies, software companies were most acquired, companies in biotechnology sector offered the most IPO and Compaq acquisition by HP was the highest acquisition in the dataset of $25 billion.

Classification: According to our results, the top three classification models which have shown the best results are XGBoost, KNN, and Random Forest. In the case of classification without sampling the dataset that is classification without balancing the imbalance data, KNN and XGBoost have shown the best performance. Out of all the under-sampling techniques used which are Random under-sampling, One sided selection, Edited nearest neighbor (ENN), Neighborhood cleaning rule, and Tomek link, ENN has shown the best performance. KNN gave the best modeling results for the ENN method with micro F1 of 0.956, macro F1 of 0.709, and 1- hamming loss of 0.956. So the KNN results outweigh the other ML algorithms and obtain the best results.

Further the optimal selection of the nearest neighbour for KNN model is also discussed and the analysis shows that the 'k' = 3 performs better than other values of the nearest neighbour. Finally the results are concluded with the analysis of the different dataset which predict the same target class. The CrunchBase dataset along with the macroeconomic variables and Intellectual Property variables shows the best performance in the terms of the micro-F1, macro-F1 and 1-hamming loss scores. Other

combinations of the dataset do not show any improvement in the results discussed earlier.

Recommendations:

- The additional variables like the degree or the educational qualification of the employees could not be incorporated into the dataset due to a lack of resources. The literature review of the M&A prediction (Ying, 2020) with the ML techniques suggest that the degree of the employees impact the M&A events and also the valuation of the firm, so future work in this domain can focus on including the degree dataset and accordingly checking the performance of the models with the literature work.

- The financial performance of the companies has impacted the M&A events in the future and the current trend (Tsagkanos, 2007) shows that they will impact the valuation, funding, and finally the M&A events in the future. So the future work can also include the financial performance and the ratio such as liquidity ratio, D/E ratio, P/R ratio among others.

- Oversampling techniques and more classifiers can be tested using good processing systems.

CHAPTER 6

LIMITATIONS OF THE STUDY

- The dataset used during the conduct of this project had data extracted till 2015, but if adequate resources were there the most recent dataset of the companies till 2021 could have been used for preprocessing and modeling purposes.

- We did not have a high computational system to process large datasets, so to model and find the results, under-sampling techniques were deployed on the dataset and then the machine learning models were applied. If a system with a good GPU or lab-based system could have been used, the imbalance dataset handling techniques such as SMOTE and other oversampling techniques could be tried and the results could be compared with those obtained via the under-sampling techniques.

- Along with under-sampling techniques, we could have also applied good classifiers like SVM which our existing system could not process.

CHAPTER 7

REFERENCES

1. Batista, F., and Carvalho, J. P. (2015, August). Text based classification of companies in CrunchBase. In 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) (pp. 1-7). IEEE.

2. Brownlee, J (2020, Aug 3), Repeated k-Fold Cross-Validation for Model Evaluation in Python. Retrieved May 02, 2021, from https://machinelearningmastery.com/repeated-k-fold-cross-validation-with-python/

3. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

4. Färber, M., Menne, C., and Harth, A. (2018). A linked data wrapper for crunchbase. Semantic Web, 9(4), 505-515.

5. Github. (2015), notpeter CrunchBase. [Data file]. Retrieved from https://github.com/notpeter/crunchbase-data

6. Hart, P. (1968). The condensed nearest neighbor rule (corresp.). IEEE transactions on information theory, 14(3), 515-516.

7. Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In Icml (Vol. 97, pp. 179-186).

8. Liu, H., Hu, Y., & Ma, W. (2011, August). Forecasting the Price of the Candidate in M&A Based on Multiple-Kernel SVMR. In *2011 International Conference on Management and Service Science* (pp. 1-4). IEEE.

9. Lee, K., Joo, S., Baik, H., Han, S., & In, J. (2020). Unbalanced data, type II error, and nonlinearity in predicting M&A failure. *Journal of Business Research*, *109*, 271-287.

10. Liang, Y. E., and Yuan, S. T. D. (2016). Predicting investor funding behavior using crunchbase social network features. Internet Research.

11. Mani, I., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets (Vol. 126). United States: ICML.

12. Nathan, M., Kemeny, T., and Almeer, B. (2017). Using Crunchbase to explore innovative ecosystems in the US and UK.

13. Rosaen, K. (2016, June 20). K-fold cross-validation. Retrieved April 01, 2021, from http://karlrosaen.com/ml/learning-log/2016-06-20/

14. Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, 6(6), 448-452

15. Tomek, I. (1976). Two Modifications of CNN. IEEE Transactions on Systems Man and Communications, 6, 769-772.

16. Tsagkanos, A., Georgopoulos, A., & Siriopoulos, C. (2007). Predicting Greek mergers and acquisitions: A new approach. International Journal of Financial Services Management, 2(4), 289–303.

17. Xiang, G., Zheng, Z., Wen, M., Hong, J. I., Rosé, C. P., and Liu, C. (2012, June). A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. In ICWSM.

18. Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3), 408-421.

19. World Bank. (2021), World Development Indicators. [DataFile]. Retrieved from https://databank.worldbank.org/source/world-development-indicators

20. Xiang, C., Weihua, C., & Ming, X. (2009, October). Coal enterprises merger and acquisition risk prediction based on support vector machine. In *2009*

Second International Conference on Intelligent Computation Technology and Automation (Vol. 4, pp. 154-157). IEEE.

21. Yang, C. S., Wei, C. P., & Chiang, Y. H. (2014). Exploiting technological indicators for effective technology merger and acquisition (M&A) predictions. *Decision Sciences*, *45*(1), 147-174.

22. Ying, Q., & He, S. (2020). Is the CEOs' financial and accounting education experience valuable? Evidence from the perspective of M&A performance. China Journal of Accounting Studies, 8(1), 35-65.