MINOR PROJECT

"Predicting and Evaluating the Popularity of Online News"

Report submitted in partial fulfilment of the requirements for the award of

Degree of Bachelor of Technology
in
Software Engineering (SE)

Under the supervision of

Ms. Kusum Lata
(Assistant Professor, CSE)

By:

Arjun Rajpal (2K14/SE/021)
Arpit Jain (2K14/SE/022)
Avinav Goel (2K14/SE/024)

To:

Department of Computer Science and Engineering

Delhi Technological University

(Formerly Delhi College of Engineering)

DECLARATION

I hereby certify that the work which is presented in the Minor Project entitled **"*Predicting and Evaluating the Popularity of Online News*"** in fulfilment of the requirement for the award of the Degree of Bachelor of Technology and submitted to the Department of Computer Engineering, Delhi Technological University (Formerly Delhi College Of Engineering), New Delhi is an authentic record of my own, carried out during a period from August 2016 to November 2016, under the supervision of **Ms. Kusum Lata, Assistant Professor, CSE** Department.
The matter presented in this report has not been submitted by me for the award of any other degree of this or any other Institute/University.

Signature

ARJUN RAJPAL 2K14/SE/021

ARPIT JAIN 2K14/SE/022

AVINAV GOEL 2K14/SE/024

ACKNOWLEDGEMENT

“The successful completion of any task would be incomplete without accomplishing the people who made it all possible and whose constant guidance and encouragement secured us the success.”

First of all, we are grateful to the Almighty for establishing us to complete this minor project.

We are grateful to **Ms. Kusum Lata, Assistant Professor** (Department of Computer Science and Engineering), Delhi Technological University (Formerly Delhi College of Engineering), New Delhi and all other faculty members of our department, for their astute guidance, constant encouragement and sincere support for this project work.

We owe a debt of gratitude to our guide, **Ms. Kusum Lata, CSE** Department for incorporating in us the idea of a creative Minor Project, helping us in undertaking this project and also for being there whenever we needed her assistance.

I also place on record, my sense of gratitude to one and all, who directly or indirectly have lent their helping hand in this venture. We feel proud and privileged in expressing my deep sense of gratitude to all those who have helped me in presenting this project.

Last but never the least, we thank our parents for always being with us, in every sense.

CERTIFICATE

This is to certify that ARJUN RAJPAL 2K14/SE/021, ARPIT JAIN 2K14/SE/022 and AVINAV GOEL 2K14/SE/024, the bonafide students of Bachelor of Technology in Software Engineering of Delhi Technological University (Formerly Delhi College Of Engineering), New Delhi of 2014–2018 batch have completed their minor project entitled "***Predicting and Evaluating the Popularity of Online News***" under the supervision of Ms. Kusum Lata, Assistant Professor, CSE DEPARTMENT.

It is further certified that the work done in this dissertation is a result of candidate's own efforts. I wish his/her all success in her life.

Date:

Ms. Kusum Lata
Assistant Professor
Computer Science & Engineering
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad, Daulatpur, Bawana Road, Delhi –110042

ABSTRACT

Consider the situation where an online news publishing agency is browsing submissions of news articles, but can only accept a few without going over budget or without overwhelming the audience. How does the agency determine which news articles will become popular or even viral, and which news articles will be ignored by the general public? Are there any predictors that indicate how many times an article will be shared amongst audiences? Implied in this question is the classification problem of binning. Our class variable, the number of shares, is a metric that defines how often an article is shared on social media, but it is a continuous variable and so its binning is not obvious. However, we are interested primarily in articles with high popularity as our positive class, more so than articles with low popularity. So how do we bin a numerical attribute into classes of ‘obscure’, ‘mediocre’, popular’, ‘viral’, etc.?

With the expansion of the Internet, more and more people enjoys reading and sharing online news articles. The number of shares under a news article indicates how popular the news is. In this project, we intend to find the best model and set of feature to predict the popularity of online news, using machine learning techniques. Our data comes from Mashable, a well-known online news website. We implemented various learning algorithms on the dataset, ranging from various regressions to SVM. Their performances are recorded and compared. Feature selection method has been used to improve performance and reduce features. SVM turns out to be the best model for prediction, and it can achieve an accuracy of 67% with optimal parameters. Our work can help online news companies to determine news popularity before publication.

TABLE OF CONTENTS

Declaration i

Acknowledgement ii

Certificate iii

Abstract iv

List of Figures ix

Chapter

1. INTRODUCTION 1

    1.1 OBJECTIVE 1

2. DATA MINING 4

    2.1 WHAT IS DATA MINING? 4

    2.2 DATA MINING FUNCTIONALITIES 5

        2.2.1 CLASSIFICATION AND PREDICTION 5

        2.2.2 CLUSTER ANALYSIS 7

        2.2.3 ARE ALL THE PATTERNS INTERESTING 7

        2.2.4 CLASSIFICATION OF DATA MINING SYSTEMS 8

3. DATA PROCESSING 10

    3.1 WHY PREPROCESS DATA? 10

    3.2 DESCRIPTIVE DATA SUMMARISATION 11

        3.2.1 MEASURING THE CENTRAL TENDENCY 11

        3.2.2 MEASURING THE DISPERSION OF DATA 13

    3.3 GRAPHIC DISPLAYS OF BASIC DESCRIPTIVE DATA SUMMARIES 16

    3.4 FORMS OF DATA PREPROCESSING 21

3.4.1 DATA CLEANING 22

3.4.2 DATA INTEGRATION 25

3.4.3 DATA TRANSFORMATION 26

3.4.4 DATA REDUCTION 27

4. MACHINE LEARNING ALGORITHMS SELECTION & ASSESSMENT 28

4.1 INTRODUCTION 28

4.2 WHAT IS CLASSIFICATION? WHAT IS PREDICTION? 30

4.3 PREPARING THE DATA FOR CLASSIFICATION AND PREDICTION 30

4.3.1 OVERFITTING AND DATA SPLITTING 30

4.4 COMPARING CLASSIFICATION AND PREDICTION METHODS 32

4.5 LINEAR REGRESSION 33

4.6 CLASSIFICATION BY REGRESSION BASED METHODS 37

4.6.1 LOGISTICS REGRESSION 38

4.7 SUPPORT VECTOR MACHINE(SVM) 40

4.7.1 LINEAR SVM 41

4.7.2 NON-LINEAR CLASSIFICATION 41

4.7.3 COMPLEXITY ANALYSIS 42

4.7.4 EVALUATION OF BINARY CLASSIFIERS 42

4.7.4.1 SENSITIVITY AND SPECIFICITY 42

4.7.4.2 POSITIVE AND NEGATIVE PREDICTIVE VALUES 42

4.7.4.3 SINGLE METRICS 43

4.8 MODEL SELECTION 43

4.8.1 ROC CURVES 43

5. PRACTICAL IMPLEMENTATION 46

5.1 DATASET ANALYSIS 46

5.2 DATA PREPROCESSING 48

5.3 NORMALIZATION 52

5.4 REMOVAL OF COLLINEAR ATTRIBUTES 53

6. MODEL SELECTION 55

6.1 MULTIVARIATE LINEAR REGRESSION 55

6.2 LOGISTIC REGRESSION 59

6.3 SUPPORT VECTOR CLASSIFICATION 65

7. RESULT AND CONCLUSION 71

7.1 RESULTS OF LINEAR REGRESSION 71

7.1.1 WITHOUT FEATURE SELECTION 71

7.1.2 WITH FEATURE SELECTION 71

7.2 RESULTS OF LOGISTICS REGRESSION 71

7.2.1 WITHOUT FEATURE SELECTION 71

7.2.2 WITH FEATURE SELECTION 72

7.3 RESULTS OF SUPPORT VECTOR CLASSIFICATION 72

7.3.1 WITHOUT FEATURE SELECTION 72

7.3.2 WITH FEATURE SELECTION 73

7.4 CONCLUSION 73

8. REFERENCES 74

9. FUTURE SCOPE 75

LIST OF FIGURES

Figure 2.1 Data Mining as a step in the process of knowledge discovery.
Figure 2.2 A classification model can be represented in various forms.
Figure 2.3 A 2-D plot, showing three data clusters.
Figure 2.4 Data mining as a confluence of multiple disciplines.
Figure 3.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.
Figure 3.2 Example of box-plot showing the unit price data for items sold at four branches of all Electronics during a given time period.
Figure 3.3 Example of histogram.
Figure 3.4 Transformation of histogram into kernel density plot.
Figure 3.5 Example of quantile plot.
Figure 3.6 Example of quantile-quantile plot.
Figure 3.7 Example of scatter plot.
Figure 3.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.
Figure 3.9 Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.
Figure 3.10 Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.
Figure 3.11 Forms of Data Preprocessing.
Figure 4.1 Training, test and validation set.
Figure 4.2 Sample data for Chemical Process.
Figure 4.3 Scatter plot of Temperature vs Yield.
Figure 4.4 Regression plane and contour plot for regression model.
Figure 4.5 Comparison of linear regression model and logistic regression model.
Figure 4.6 Example of Linear Classifier
Figure 4.7 ROC Space
Figure 4.8 The ROC curves of two classification models.
Figure 5.1 Comparison of Actual vs Recorded Data for an article(leaked: More Low Cost iPhone photos)
Figure 5.2 Frequency Distribution histogram showing outlier observations
Figure 5.3 Box-plot of Original data attributes
Figure 5.4 Frequency Distribution after Box Plot
Figure 5.5 Mean, Standard Deviation and Normalization Formula

Figure 6.1 Minimization of Cost Function in Gradient Descent with Feature Selection
Figure 6.2 Minimization of Cost Function in Gradient Descent without Feature Selection
Figure 6.3 ROC Curve for Logistic Classifier without Feature Selection
Figure 6.4 ROC Curve for Logistic Classifier with Feature Selection
Figure 6.5 ROC Curve for SVM Classifier with Feature Selection
Figure 6.6 ROC Curve for SVM Classifier without Feature Selection