# **SLC FINAL EXAM PAPER FOR 70 MARKS**

# Data Set Information:

* The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content be publicly accessed and retrieved using the provided urls.
* Acquisition date: January 8, 2015
* The estimated relative performance values were estimated by the authors using a Random Forest classifier and a rolling windows as assessment method. See their article for more details on how the relative performance values were set.

# Attribute Information:

* Number of Attributes: 61 (58 predictive attributes, 2 non-predictive, 1 goal field)

|  |  |
| --- | --- |
| **Column Names** | **description** |
| 0. url: | URL of the article (non-predictive) |
| 1. timedelta: | Days between the article publication and the dataset acquisition (non-predictive) |
| 2. n\_tokens\_title: | Number of words in the title |
| 3. n\_tokens\_content: | Number of words in the content |
| 4. n\_unique\_tokens: | Rate of unique words in the content |
| 5. n\_non\_stop\_words: | Rate of non-stop words in the content |
| 6. n\_non\_stop\_unique\_tokens: | Rate of unique non-stop words in the content |
| 7. num\_hrefs: | Number of links |
| 8. num\_self\_hrefs: | Number of links to other articles published by Mashable |
| 9. num\_imgs: | Number of images |
| 10. num\_videos: | Number of videos |
| 11. average\_token\_length: | Average length of the words in the content |
| 12. num\_keywords: | Number of keywords in the metadata |
| 13. data\_channel\_is\_lifestyle: | Is data channel 'Lifestyle'? |
| 14. data\_channel\_is\_entertainment: | Is data channel 'Entertainment'? |
| 15. data\_channel\_is\_bus: | Is data channel 'Business'? |
| 16. data\_channel\_is\_socmed: | Is data channel 'Social Media'? |
| 17. data\_channel\_is\_tech: | Is data channel 'Tech'? |
| 18. data\_channel\_is\_world: | Is data channel 'World'? |
| 19. kw\_min\_min: | Worst keyword (min. shares) |
| 20. kw\_max\_min: | Worst keyword (max. shares) |
| 21. kw\_avg\_min: | Worst keyword (avg. shares) |
| 22. kw\_min\_max: | Best keyword (min. shares) |
| 23. kw\_max\_max: | Best keyword (max. shares) |
| 24. kw\_avg\_max: | Best keyword (avg. shares) |
| 25. kw\_min\_avg: | Avg. keyword (min. shares) |
| 26. kw\_max\_avg: | Avg. keyword (max. shares) |
| 27. kw\_avg\_avg: | Avg. keyword (avg. shares) |
| 28. self\_reference\_min\_shares: | Min. shares of referenced articles in Mashable |
| 29. self\_reference\_max\_shares: | Max. shares of referenced articles in Mashable |
| 30. self\_reference\_avg\_sharess: | Avg. shares of referenced articles in Mashable |
| 31. weekday\_is\_monday: | Was the article published on a Monday? |
| 32. weekday\_is\_tuesday: | Was the article published on a Tuesday? |
| 33. weekday\_is\_wednesday: | Was the article published on a Wednesday? |
| 34. weekday\_is\_thursday: | Was the article published on a Thursday? |
| 35. weekday\_is\_friday: | Was the article published on a Friday? |
| 36. weekday\_is\_saturday: | Was the article published on a Saturday? |
| 37. weekday\_is\_sunday: | Was the article published on a Sunday? |
| 38. is\_weekend: | Was the article published on the weekend? |
| 39. LDA\_00: | Closeness to LDA topic 0 |
| 40. LDA\_01: | Closeness to LDA topic 1 |
| 41. LDA\_02: | Closeness to LDA topic 2 |
| 42. LDA\_03: | Closeness to LDA topic 3 |
| 43. LDA\_04: | Closeness to LDA topic 4 |
| 44. global\_subjectivity: | Text subjectivity |
| 45. global\_sentiment\_polarity: | Text sentiment polarity |
| 46. global\_rate\_positive\_words: | Rate of positive words in the content |
| 47. global\_rate\_negative\_words: | Rate of negative words in the content |
| 48. rate\_positive\_words: | Rate of positive words among non-neutral tokens |
| 49. rate\_negative\_words: | Rate of negative words among non-neutral tokens |
| 50. avg\_positive\_polarity: | Avg. polarity of positive words |
| 51. min\_positive\_polarity: | Min. polarity of positive words |
| 52. max\_positive\_polarity: | Max. polarity of positive words |
| 53. avg\_negative\_polarity: | Avg. polarity of negative words |
| 54. min\_negative\_polarity: | Min. polarity of negative words |
| 55. max\_negative\_polarity: | Max. polarity of negative words |
| 56. title\_subjectivity: Title subjectivity | Title subjectivity |
| 57. title\_sentiment\_polarity: | Title polarity |
| 58. abs\_title\_subjectivity: | Absolute subjectivity level |
| 59. abs\_title\_sentiment\_polarity: | Absolute polarity level |
| 60. shares: | Number of shares (target) |

* **Instructions:**  Based on the above dataset the students are expected to follow the below mentioned steps.

|  |  |  |
| --- | --- | --- |
| STEPS | Instructions | MARKS |
| 1. Data Ingestion: | Read the dataset (tab, csv, xls, txt, inbuilt dataset). Do the descriptive statistics and do null value condition check, write an inference on it**.** | 5 |
| 2. Split Data set: | (train – 70% and test – 30%), Using test and train data, calculate R square value and check if the model Over-fit / under-fit model.  Write a detailed inference on it. | 5 |
| 3. Build a base model: | Build a base model (which will include all the parameter as input) model using Logistic regression considering all the input parameters, calculate all the relevant metrics and write summary | 5 |
| 4. EDA - Plots: | 1. Check for Outlier using Box plot and apply the formula to check how many of these data has outliers 2. Apply correlation plot and display with colour code and numerical indication 3. Plot best X with Y and find how the Y is varying with X. Like if X is continuous, go for scatter plot. if X is categorical, go for box plot. 4. Apply density plot and check if the data is normally distributed | 5 |
| 5. EDA - Inference: | Write detailed inference (not observations) about the question 4, few pointers like Outlier detection, outlier treatment, conditions for linearity, impact because of transformation on variables, distributions etc. | 5 |
| 6. Label Encoding / Get Dummies: | Build a base model (Include all input parameters) using logistic regression if data is binary / or use any other modelling technique if data in multiclass. | 5 |
| 7. Data handling: | Check conditions for data imputation and data imbalance for the given dataset. Handling data imputation if needed and balance the data (Either go for under sampling or smote). If not, then explain why you do not need sampling. | 10 |
| 8. Feature selection and Feature engineering | Build multiple models by using feature selection and feature engineering and identify which is the good model and write the inference | 10 |
| 9. logistic, decision tree, KNN, Naïve Bayes | Build a classification model (**logistic, decision tree, KNN, Naïve Bayes**) and calculate the metric and write inference on all the model and describe which model is very good | 10 |
| 10. Ensemble techniques: | Apply ensemble techniques (at least 2) to build model and write inference on the same |  |
| 11. Output | Display the output in tabular format, which will compare all the model, model tuning methods, metrics.  Compare all the models create and write detailed summary on the **Final** **model** and why is this best model, enrich with the factor why it is best model based on what? (e.g. EDA inference, transformation, feature selection etc.) | 10 |