STATS 202: Data Mining and Analysis

Instructor: Linh Tran

FINAL PROJECT

Kaggle submission *due date: August 12, 2024*

Write-up due date: August 14, 2024

Stanford University

Introduction

The goal of this class project is to give you experience in real life statistical analyses and data mining. By the end of the project, you will have learned how to identify and interpret types of different attributes in a dataset, visualize the attributes and relationships between attributes of different types, understand how those relationships could affect your models and analyses, and finally build regression and classification models.

The class project is optional. It can replace your score in the Final Exam if you choose not to take it. If you choose to also take the Final Exam, your score for the Final Exam portion of your grade will be the maximum of the Final Exam and the Class Project.

Background

The goal of a search engine is to return relevant documents for search queries that users enter. Search engines typically use hundreds of signals to determine the relevance of a document and then return a list of documents in order of relevance.

You will be provided a training data set which includes 10 attributes and 80,046 observations from search engine query and url data. The dataset contains 10 different signals that could be used to help predict whether the url is relevant for the query. Additionally, a test data set is provided which contains 30,001 observations. Your goal is to make relevance predictions for each row (urls for a query) in the test data set.

Task

Your job is to create a text file containing one line per example in the test set. On each line, give your prediction for the relevance (1 for relevant, 0 for not relevant) of that row in the test set. Submissions will be made to Kaggle.

Your submission will be scored according to your accuracy, i.e.

$$\text{score} = 1 - \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i) \tag{1}$$

Write-up

As part of the final project, you are expected to submit a final report (in PDF format) covering the approaches, details, and results of each of the four objectives. The report should be no longer than 5 pages (excluding figures, tables, and code) and capture the steps you took throughout the data mining process (Figure 1). Table 1 provides further details on each step of the process. Explain the decisions you made and provide visualizations supporting those decisions. Furthermore, provide visualizations in the form of tables and/or figures for each attribute in your model, and provide visualizations for pairs of attributes that you think may be related. Remember, understanding your data is an important part of the data mining

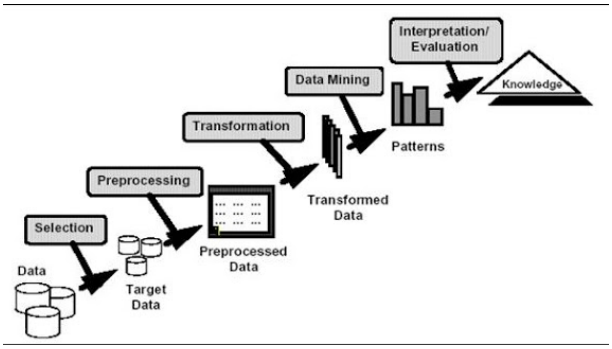

Figure 1: Steps in the data mining process.

Selection	Explain which attributes you used to build the model, and why you chose those attributes
Preprocessing	Explain whether you pre-processed any of the attributes by modifying them in any way
Transformation	Explain whether you created new features from the existing attributes, or from pairs of the existing attributes. Did you transform any of the attributes into another representation of data? Remember, you do not need to use all of the attributes in your model. Try to evaluate which attributes you think will be useful, and use those attributes.
Data Mining	Explain how you built your regression / classification model. There are many kinds of models that may work for this problem. You are welcome to use whatever regression / classification approach you would like, but remember, you need to end up with a prediction or probability.
Interpretation/Evaluation	Understand what your model is doing and how it is performing. This may require you to separate your training data into different groups so that you can test your models performance on a “hold out” group.

Table 1: Steps in the data mining process.

process, and visualization that data can help understand it. **Make sure to reference your team’s Kaggle leaderboard name in your report**.

The code used to generate the results should be either attached as additional scripts in your final project submission, appended to the end of your report as an appendix, or referenced to in the report via a link to the uploaded git repository. Note that 10% of your grade will be based upon the organization, readability, reproducibility, and efficiency of your code.