Chap3_LuRuiqi.docx

by LU RUI QI

Submission date: 24-Jul-2024 12:47PM (UTC-0700)

Submission ID: 2421918635

File name: Chap3_LuRuiqi.docx (348.8K)

Word count: 610

Character count: 3650

Chapter 3: Methodology

Data Science Project Life Cycle

The Illustration of the data science project life cycle are:

1. Problem Definition: Clearly state the problem you want to solve, for instance, the research problem investigated in the study is: predicting traffic jam in Malaysia using machine learning algorithms.

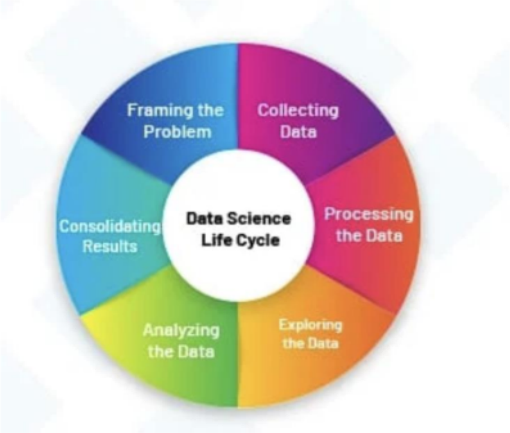
2. Data Collection: Aggregate data needed from multiple resources, such as traffic reports, weather conditions, road constructions, and public events.

3. Data Preprocessing: Clean the collected data, by imputing any missing or null values as well converting different time formats of the data and incorporating new derived features. 4. Exploratory Data Analysis (EDA): Comprehensively visualize and analyze data using graphical visualization libraries which will present important patterns, trends, and insights. 5. Model Development: Processing and training of machine learning models to gain insight and learn the complex patterns in the dataset.

6. Model Evaluation: Assessment and comparison of models against an evaluation metric to decide what model has the best performance.

7. Deployment and Monitoring: Final stage that is implementing the best performing model and monitoring its behavior in a production environment.

8. Reporting and Visualization: Interpret model results with visualizations and reports with practical interpretations of the model predictions, uncertainty and stability of the predictions.



Data Sources and Collection Methods

1. Traffic Volume Data: It's the data get from traffic sensors, cameras, or data found online from government transportation agencies.

2. Weather Data: The data is acquired from an external API like OpenWeatherMap, weather source website providing historical weather data.

3. Public Transportation Ridership Data: Data can be collected from public transportation agencies or through any open data website related to transportation.

4. Historical Accident Data: The data here can be gathered from, police reports or crash data reports from governing bodies or transportation related agencies.

Data Pre-processing

This stage is crucial as it ensures none of the data used is dirty or flawed before further analysis, and the involvement of data preprocessing consist of the following:

1.Data Cleaning: - Handling Missing Values: Null values can be filled in by several methods and techniques. Like: imputing mean, median, or mode for numerical data, and the use of the most frequent category for categorical. - Outliers Removal: To detect and remove outliers, statistical methods such as Z-score and IQR should be used as outliers can have a major impact on data distribution.

2.Data Transformation: - Normalization: Scale numerical features to a standard range, typically between 0 and 1, after which all features will be equally contributing towards model learning. Techniques: Min-Max Scaling, Standard Scaler. - Encoding Categorical Variables: Transformation of categorical data to numerical quantities. Techniques: One-Hot Encoding, Label encoder.

3. Feature Engineering: Creating New Features: Generation of other features from given data. Examples: time of day, traffic density, road occupancy, time-travel distance impact, and the cumulative weather scores. - Feature Selection: Choose which relevant features to use. Methods: correlation, Mutual Information score, feature importance scores from Random Forest, etc.

First Exploratory Data Analysis and Results

1. Traffic Volume Over Time: Line plots of how the traffic volume changes every hour and every day of the week.

2. Heatmaps: Correlation matrices of various features and their relationship with the traffic volume.

3. Descriptive Statistics: Mean, median, standard deviation, and range for the traffic volume and other numerical features.

Initial Machine Learning Models

1. Linear Regression: As a naive model to predict the traffic volume. Performance Metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

2. Random Forest: Consider non-linear interactions between different predictors. Performance Metrics: MAE, RMSE, R-squared.

3. Gradient Boosting: Ensemble method combining multiple weak learners which combines things well to improve overall performance. Performance Metrics: MAE, RMSE, R-squared.

Chap3_LuRuiqi.docx

ORIGINALITY REPORT

4%	4%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ijates.com Internet Source	3%
2	etd.cput.ac.za Internet Source	2%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		