Optimizing Predictive Models for Water Tanker Demand: A Comparative Analysis

A PROJECT REPORT

Submitted by

PUNEETH H S & KISHOR DESAI

in partial fulfillment for the award of the degree

of

BSC (HONS.) C.S.

School of Computer Science and Engineering
RV University
RV Vidyaniketan,8th Mile, Mysuru Road, Bengaluru, Karnataka, India - 562112

JULY 2024

TABLE OF CONTENTS

ABSTRACT

The increasing challenges of urban water management necessitate effective resource allocation strategies, particularly in densely populated cities like Bangalore, India. Predicting water tanker demand is vital to addressing urban water shortages and optimizing resource distribution. This study explores the development and optimization of predictive models to forecast water tanker demand accurately, focusing on addressing the impact of data outliers on model performance.

Two machine learning models were implemented for the task: Linear Regression and Random Forest Regression. The primary aim was to improve the predictive accuracy of these models through effective data preprocessing techniques, particularly the removal of outliers. Data preprocessing included essential steps such as cleaning columns and rows with missing values, typecasting numerical data stored in incorrect formats, and applying One-Hot Encoding to categorical variables. The significance of preprocessing was evident in the substantial performance improvement of both models after addressing data inconsistencies.

The Linear Regression model's accuracy improved significantly, rising from 0.01 to 0.74 after preprocessing and outlier removal. Similarly, the Random Forest Regression model showed notable improvement, with its accuracy increasing from -0.15 to 0.75. These findings underscore the critical role of outlier removal in enhancing the quality of the dataset and improving model reliability. Furthermore, the study highlights the differences in how Linear Regression and Random Forest Regression respond to preprocessing, offering practical insights for urban water management.

This study contributes to the field by providing a comparative analysis of model performance before and after data preprocessing, emphasizing the importance of robust data cleaning techniques. While the study is limited by the scope of its dataset and urban context, the results demonstrate the potential of applying such methods to similar urban planning scenarios. Future research could explore advanced modeling techniques and integrate diverse datasets to further improve prediction accuracy. The findings of this study have practical implications for urban planners and data scientists, enabling more efficient and data-driven resource allocation strategies in water tanker distribution.

1. INTRODUCTION

1.1 Background

Urban water management is a critical challenge in rapidly growing cities, especially in developing countries like India. Cities such as Bangalore face acute water shortages due to increasing population demands and limited resources. Water tanker distribution plays a vital role in addressing these shortages, particularly during peak demand periods. However, without accurate forecasting of water tanker demand, resource allocation becomes inefficient, leading to wasted resources and unmet needs.

Predictive modeling, driven by machine learning techniques, offers a solution to this problem. By analyzing historical data, these models can predict water tanker demand, allowing for better planning and distribution. However, issues such as data inconsistencies, missing values, and outliers limit the effectiveness of predictive models, resulting in inaccurate forecasts. Addressing these challenges is essential for improving prediction accuracy and enhancing urban water management.

1.2 Objectives

This study aims to develop and optimize predictive models to forecast water tanker demand. It focuses on:

- Improving data preprocessing methods to address inconsistencies and outliers.
- Comparing the performance of Linear Regression and Random Forest Regression models.
- Providing actionable insights for urban planners to enhance resource allocation strategies.

1.3 Methodology Overview

The study employs a systematic approach to address data quality issues and evaluate predictive models. Key steps include:

1. Data preprocessing to clean and prepare the dataset.
2. Implementation of Linear Regression and Random Forest Regression models.
3. Evaluation of model performance before and after data preprocessing using accuracy, RMSE, and MAE.

By addressing data-related challenges and leveraging machine learning models, this study contributes to improving urban water management systems and ensuring efficient resource distribution.

2. Related work

Numerous studies have explored the use of machine learning techniques for demand forecasting in urban planning and resource management. However, significant gaps remain in addressing data quality issues, such as outliers and inconsistencies, which often reduce model accuracy and reliability.

For instance, Xiang et al. (2020) demonstrated the application of artificial intelligence techniques in urban water resource management, emphasizing the importance of accurate forecasting for sustainable planning. However, their study did not focus on preprocessing techniques to handle data inconsistencies, which limited the robustness of their models.
Pacchin et al. (2019) compared short-term water demand forecasting models but highlighted the challenges posed by outliers in the datasets. While their work provided insights into the selection of models, it lacked a systematic approach to outlier removal, which could have improved model performance.
Herrera et al. (2010) developed predictive models for hourly urban water demand but relied heavily on raw datasets without accounting for extreme values or missing data. This reliance led to inaccuracies in predictions, particularly during periods of irregular demand.

The current study builds on this existing research by addressing these critical gaps. Unlike prior work, it implements a comprehensive preprocessing strategy, including outlier removal, typecasting, and encoding, to improve data quality before training the models. Additionally, it compares the performance of Linear Regression and Random Forest Regression models to evaluate their robustness post-preprocessing.

By demonstrating significant improvements in model accuracy (from 0.01 to 0.74 for Linear Regression and from -0.15 to 0.75 for Random Forest Regression), this study highlights the effectiveness of preprocessing techniques in overcoming the limitations of previous work. The findings not only improve predictive accuracy but also provide a replicable framework for urban water management applications.
Future research could extend this work by incorporating advanced modeling techniques, such as ensemble learning or neural networks, to further enhance predictive performance in diverse urban contexts.

3. METHODOLOGY

3.1 Methodological Approach

The methodological approach of this study focuses on developing predictive models for forecasting water tanker demand using Linear Regression and Random Forest Regression. The process emphasizes the importance of preprocessing data to improve model performance, addressing issues such as missing values, inconsistent formats, and outliers.

This approach involves:

1. Cleaning and preprocessing data to ensure quality and consistency.
2. Implementing and training machine learning models.
3. Evaluating the models using appropriate metrics to measure their performance.

The comparative analysis of model performance before and after preprocessing provides insights into the impact of data quality on predictive accuracy.

3.2 Data Collection and Selection

The dataset used in this study includes monthly records of water tanker orders, their associated costs, and related variables from an urban area in India.

- **Data Source**: Historical records provided by water management authorities.
- **Data Features**: Includes categorical and numerical variables such as the number of tankers ordered, total cost spent, and operational conditions.
- **Selection Criteria**:
    - Only records with sufficient and meaningful data were included.
    - Rows with excessive missing values were excluded to maintain data integrity.

The selected dataset represents a real-world scenario of water tanker demand in urban areas, making it suitable for the study's objectives.

3.3 Methods of Data Analysis

The study employs the following data preprocessing and modeling techniques:

1 Data Preprocessing

- **Column and Row Cleaning**:
    - Removed empty columns and rows with missing data that could not be imputed.
    - Justification: Ensures that incomplete or irrelevant data does not bias the model.
- **Typecasting**:
    - Converted numerical values stored as objects into the correct data types.

  - Justification: Prevents errors during model training and improves reliability of numerical operations.

- **Encoding**:
  - Applied One-Hot Encoding to categorical variables.
  - Justification: Allows categorical data to be effectively utilized in the regression models.
- **Outlier Removal**:
  - Detected and removed extreme data points that skewed the dataset.
  - Justification: Outliers mislead the models, reducing prediction accuracy.

2 Model Implementation

- **Linear Regression**:
  - Chosen for its simplicity and interpretability.
  - Used as a baseline model to compare performance improvements after preprocessing.
- **Random Forest Regression**:
  - Selected for its robustness and ability to handle non-linear relationships in data.

Both models were trained on the cleaned dataset and evaluated based on their predictive accuracy, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error).

3.4 Evaluation and Justification of Methodological Choices

- **Preprocessing Methods**:
  Data preprocessing techniques were chosen to address common issues in real-world datasets, such as missing values and outliers. These methods are widely recognized for their effectiveness in improving model reliability and accuracy.
- **Model Selection**:
  Linear Regression was selected for its simplicity and ease of interpretation, serving as a benchmark model. Random Forest Regression was chosen for its ability to capture complex data interactions and handle non-linear patterns.
- **Evaluation Metrics**:
  RMSE and MAE were used to evaluate the models, as they provide a detailed understanding of prediction errors, complementing the accuracy score.

By combining effective preprocessing techniques with robust evaluation metrics, this methodology ensures reliable and actionable results for water tanker demand forecasting.

IMPLEMENTATION

The implementation of this project is based on the development and optimization of predictive models for water tanker demand forecasting. Below is the experimental procedure, which can be replicated to achieve the same results.

4.1 Dataset Preparation

1. **Data Collection**:
   - Gather the historical water tanker demand data, including the number of tankers ordered, total costs, and relevant features such as date, season, and weather conditions.
   - Ensure that the dataset is comprehensive and covers multiple months or years for better predictive accuracy.
2. **Data Cleaning**:
   - **Remove Empty Columns**: Identify and delete any columns with no data to prevent unnecessary complexity.
   - **Handle Missing Values**: Drop rows that have significant missing values that cannot be imputed meaningfully, or impute the missing data using appropriate techniques (mean imputation for numerical data, mode imputation for categorical data).
   - **Remove Duplicates**: Check and remove any duplicate rows in the dataset that might bias the model's predictions.
3. **Typecasting**:
   - Ensure that all numerical columns are correctly typed as integers or floats and that categorical columns are converted to an appropriate format. For example, convert the 'Date' column to a datetime type if required.
   - This step ensures that the data is in the proper format for analysis and model building.
4. **Encoding Categorical Variables**:
   - Apply **One-Hot Encoding** to any categorical columns (e.g., season, day of the week) to convert them into binary numerical variables.
   - This prepares categorical data for use in machine learning models like Linear Regression and Random Forest Regression.
5. **Outlier Detection and Removal**:
   - Identify outliers in the numerical columns (e.g., total cost or number of tankers ordered) using techniques like the IQR (Interquartile Range) or Z-scores.
   - Remove or adjust data points identified as outliers to improve model accuracy.

4.2 Model Implementation

1. **Split Data**:
    - Split the dataset into training and testing sets (e.g., 80% for training and 20% for testing).
    - This ensures that the model is trained on one set of data and evaluated on a separate set to prevent overfitting.
2. **Linear Regression Model**:
    - **Model Initialization**: Import the Linear Regression model from a machine learning library (e.g., Scikit-learn in Python).
    - **Training**: Train the Linear Regression model on the training dataset.
    - **Evaluation**: Use the testing dataset to evaluate the model's performance. Record metrics such as accuracy, RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error).
3. **Random Forest Regression Model**:
    - **Model Initialization**: Import the Random Forest Regression model from the machine learning library.
    - **Training**: Train the Random Forest model on the same training dataset.
    - **Evaluation**: Use the testing dataset to evaluate the Random Forest model's performance, recording the same evaluation metrics (accuracy, RMSE, MAE).

4.3 Model Evaluation

1. **Accuracy Evaluation**:
    - Compare the predicted values of both models with the actual values from the testing dataset.
    - Calculate the accuracy of the models based on the proportion of correct predictions.
2. **RMSE and MAE**:
    - Calculate the **Root Mean Squared Error (RMSE)** and **Mean Absolute Error (MAE)** for both models to understand the magnitude of prediction errors.
    - These metrics provide insights into how well the models are performing and whether one model is superior to the other.
3. **Comparison of Results**:
    - Compare the performance of Linear Regression and Random Forest Regression before and after outlier removal.
    - Analyze improvements in accuracy, RMSE, and MAE post-preprocessing to assess the impact of data cleaning and outlier removal.

4. RESULT AND DISCUSSION

5.1 Results

The results of the predictive models, **Linear Regression** and **Random Forest Regression**, before and after outlier removal, are presented below:

Target: Total Cost Spent on Water Tanker in a Month

- **Linear Regression**:
  - Initial accuracy: **0.01**
  - Post-outlier removal accuracy: **0.74**
- **Random Forest Regression**:
  - Initial accuracy: **-0.15**
  - Post-outlier removal accuracy: **0.75**

Target: Number of Tankers Ordered in a Month

- **Linear Regression**:
  - Post-outlier removal accuracy: **0.37**
- **Random Forest Regression**:
  - Post-outlier removal accuracy: **0.16**

Both models show significant improvements in accuracy after outlier removal, particularly for the "Total Cost Spent" target. The **Random Forest Regression** model, which initially showed negative accuracy due to being misled by outliers, demonstrates a substantial recovery after outlier removal.

5.2 Interpretation and Explanation of Results

The **Linear Regression** model demonstrated an increase in accuracy from **0.01** to **0.74** for the target "Total Cost Spent on Water Tanker in a Month," reflecting the model's ability to improve its predictions once the outliers were removed. **Random Forest Regression** also showed an improvement from a negative value (**-0.15**) to **0.75**, underscoring the model's sensitivity to outliers in the initial dataset. This result highlights the importance of preprocessing, especially outlier removal, for improving predictive accuracy in machine learning models.

However, for the target "Number of Tankers Ordered in a Month," the **Linear Regression** model performed better (**0.37**) than the **Random Forest Regression** model (**0.16**). Despite the outlier removal, the **Random Forest Regression** model did not perform as well in this specific case, suggesting that this model might require further tuning or a more complex dataset to handle the interactions in this target variable.

These results indicate that while both models benefit from outlier removal, the improvements vary based on the target variable. For cost-related predictions, both models improve

significantly, but the number of tankers ordered presents more challenges, especially for **Random Forest Regression**.

5.3 Justification of Approach

The decision to use both **Linear Regression** and **Random Forest Regression** was justified, as it allowed for a comparative analysis of simple versus more complex models. **Linear Regression** was chosen for its interpretability and simplicity, while **Random Forest Regression** was selected for its ability to capture non-linear relationships and interactions between features. The combination of these models provided a well-rounded approach to understanding predictive accuracy in water tanker demand forecasting.

The use of **outlier removal** as a preprocessing step was also justified, as it directly addressed the impact of extreme data points on model performance. By eliminating outliers, we allowed both models to focus on the core data patterns, improving their ability to make accurate predictions.

5.4 Critical Evaluation of the Study

While the study demonstrates the importance of data preprocessing, there are several limitations that should be considered:

1. **Data Quality**:
   The quality of the dataset plays a significant role in model performance. While preprocessing steps improved the models' accuracy, the dataset's inherent quality (e.g., missing values, noise, or inaccurate records) may still limit the overall performance. Future research should focus on improving data collection processes to ensure high-quality data.
2. **Model Complexity**:
   The **Random Forest Regression** model showed less improvement than expected in the target "Number of Tankers Ordered in a Month." This indicates that further tuning, such as hyperparameter optimization or exploring more complex models (e.g., ensemble methods or neural networks), could improve performance. Additionally, experimenting with other feature engineering techniques may help capture deeper interactions in the data.
3. **Generalizability**:
   The study is based on data from a specific urban area (e.g., Bangalore). While the results are valuable for this context, their applicability to other urban areas or regions with different water supply systems might be limited. Future research could expand the study by including data from other regions to validate the generalizability of the findings.
4. **Impact of Outlier Removal**:
   The significant improvement in model accuracy after outlier removal demonstrates the importance of preprocessing, but it also raises questions about the extent to which data

points should be removed. There is a balance to strike between removing outliers and retaining enough data to capture variability in real-world scenarios. More nuanced methods of outlier handling, such as outlier detection algorithms or trimming techniques, could be explored in future work.

5.5 Relation to Literature Review

This study's findings align with existing research that highlights the importance of data preprocessing in predictive modeling, particularly for forecasting demand in urban settings. Previous studies have indicated that preprocessing can significantly improve model performance by reducing the noise caused by inconsistencies and outliers. This work further validates the effectiveness of outlier removal in enhancing predictive accuracy, particularly in the context of water tanker demand.

However, our study extends this understanding by providing a direct comparison between **Linear Regression** and **Random Forest Regression** in the context of outlier removal, offering insights into which models are more sensitive to data quality. This contribution is valuable for urban planners and data scientists in selecting the right model for similar predictive tasks.

5. CONCLUSION

Summary of Findings

This study aimed to develop predictive models for forecasting water tanker demand using machine learning techniques, specifically **Linear Regression** and **Random Forest Regression**. Through extensive data preprocessing, including outlier removal, we observed significant improvements in model performance. For the target variable "Total Cost Spent on Water Tanker in a Month," both models showed substantial accuracy gains, with **Linear Regression** improving from **0.01** to **0.74** and **Random Forest Regression** improving from **-0.15** to **0.75** after outlier removal. For the target "Number of Tankers Ordered in a Month," **Linear Regression** performed better with an accuracy of **0.37**, while **Random Forest Regression** showed a modest accuracy of **0.16**.

Significance of the Project

The project highlights the importance of data preprocessing in machine learning model development, particularly in the context of urban water management. The removal of outliers was essential in improving model accuracy and ensuring more reliable predictions. By comparing two regression models, we have gained valuable insights into which models are more sensitive to data quality issues and how they perform under different preprocessing strategies. This research demonstrates that even simple models like **Linear Regression** can be highly effective when the data is properly cleaned and prepared.

Achievements

The primary achievement of this study lies in demonstrating the impact of **outlier removal** on the predictive accuracy of both **Linear Regression** and **Random Forest Regression** models. By improving the quality of the dataset, we were able to significantly enhance model performance, particularly for the prediction of **Total Cost Spent** on water tankers. This achievement can lead to more accurate forecasting, which is crucial for urban water management systems, enabling more efficient resource allocation and planning.

Moreover, this research provides a comparative analysis of two widely used regression models, contributing to a better understanding of their strengths and limitations when applied to urban water demand prediction. The findings offer actionable insights that can guide future research in predictive modeling for urban planning and resource management.

Recommendations

Based on the findings of this study, the following recommendations can be made:

1. **Further Exploration of Model Optimization**: While **Random Forest Regression** showed improvements, its performance in predicting the number of tankers ordered could be enhanced through further model optimization. Hyperparameter tuning and exploring more advanced machine learning techniques, such as ensemble methods or deep learning models, could provide additional accuracy gains.
2. **Inclusion of More Diverse Datasets**: The study was conducted with data from a single urban area. To improve the generalizability of the findings, future research should include

data from multiple regions or cities, each with its unique water supply challenges. This would help validate the applicability of the models and preprocessing techniques across different contexts.

3. **Advanced Preprocessing Techniques**: While outlier removal significantly improved the model performance, other preprocessing methods, such as advanced imputation techniques for missing values or feature engineering strategies, could be explored to further refine the models.
4. **Integration of Real-Time Data**: Future models could incorporate real-time data feeds (e.g., weather data, population growth, or water consumption trends) to improve predictions and better adapt to changing urban conditions. This would lead to even more dynamic and responsive water tanker demand forecasting systems.
5. **Collaboration with Urban Planners**: The insights gained from this study can be used by urban planners and data scientists to improve water distribution systems and optimize tanker usage. Collaboration between data scientists, urban planners, and policymakers could ensure that predictive models are integrated effectively into decision-making processes for more sustainable water resource management.

Final Thoughts

In conclusion, this study emphasizes the critical role of data preprocessing and the careful selection of machine learning models in solving real-world urban problems. The improvements in predictive accuracy after addressing data quality issues suggest that machine learning can be a powerful tool for enhancing urban water management. Future research and model development should continue to explore new methods to refine predictions and contribute to the development of smarter, more efficient urban resource management strategies.

6. FUTURE SCOPE

In the future, the model could be enhanced by integrating more diverse datasets from different urban areas to improve generalizability. Advanced techniques such as ensemble methods and deep learning could be explored for better prediction accuracy. Additionally, incorporating real-time data could help make the model more responsive to dynamic changes in water tanker demand. Further research could also focus on optimizing data preprocessing methods to handle more complex data challenges.

REFERENCES

1. **Xiaojun Xiang, Qiong Li, Shahnawaz Khan, Osamah Ibrahim Khalaf** (2020). "Urban Water Resource Management for Sustainable Environment Planning Using Artificial Intelligence Techniques." *Journal of Environmental Management*.
2. **E. Pacchin, F. Gagliardi, S. Alvisi, M. Franchini** (2019). "A Comparison of Short-Term Water Demand Forecasting Models." *Water Resources Management*, 33(3), 587-604.
3. **Michelle Sapitang, Wanie M. Ridwan, Khairul Faizal Kushiar, Ali Najah Ahmed, Ahmed El-Shafie** (2020). "Machine Learning Application in Reservoir Water Level Forecasting for Sustainable Hydropower Generation Strategy." *Environmental Monitoring and Assessment*, 192(8), 496.
4. **Siva Rama Krishnan, M. K. Nallakaruppan, Rajeswari Chengoden, Srinivas Koppu, M. Iyapparaja, Jayakumar Sadhasivam, Sankaran Sethuraman** (2022). "Smart Water Resource Management Using Artificial Intelligence—A Review." *Journal of Environmental Science and Technology*, 56(10), 2123-2140.
5. **Yanping Wang, Saeid Razmjooy** (2024). "Prediction of Drought Hydrological and Water Scarcity Based on Optimal Artificial Intelligence by Developing a Metaheuristic Optimization Algorithm." *Water Resources Research*, 60(1), 43-60.
6. **Ze Liu, Jingzhao Zhou, Xiaoyang Yang, Zechuan Zhao, Yang Lv** (2024). "Research on Water Resource Modeling Based on Machine Learning Technologies." *International Journal of Water Resources Development*, 40(5), 759-772.
7. **Guangtao Fu, Siao Sun, Lan Hoang, Zhiguo Yuan, David Butler** (2023). "Artificial Intelligence Underpins Urban Water Infrastructure of the Future: A Holistic Perspective." *Environmental Modelling & Software*, 147, 105388.
8. **Manmeet Singh, Suhaib Ahmed** (2021). "IoT-Based Smart Water Management Systems: A Systematic Review." *Sensors*, 21(6), 2039.
9. **Matthew Lowe, Ruwen Qin, Xinwei Mao** (2022). "A Review on Machine Learning, Artificial Intelligence, and Smart Technology in Water Treatment and Monitoring." *Science of the Total Environment*, 794, 148634.
10. **M. Kavya, Aneesh Mathew, Padala Raja Shekar, Sarwesh P** (2023). "Short Term Water Demand Forecast Modelling Using Artificial Intelligence for Smart Water Management." *Journal of Water Resources Planning and Management*, 149(7), 04023042.
11. **Manuel Herrera, Luís Torgo, Joaquín Izquierdo, Rafael Pérez-García** (2010). "Predictive Models for Forecasting Hourly Urban Water Demand." *Urban Water Journal*, 7(1), 1-10.
12. **Ahmed Abdel Nasser, Magdi Z. Rashad, Sherif E. Hussein** (2020). "A Two-Layer Water Demand Prediction System in Urban Areas Based on Micro-Services and LSTM Neural Networks." *Journal of Water Resources Management*, 34(4), 1151-1164.