PAPER NAME

Enhancing Property Selling with Predictive Analytics using Random Forest

AUTHOR

Anandhu Biju

WORD COUNT

2939 Words

CHARACTER COUNT

16987 Characters

PAGE COUNT

5 Pages

FILE SIZE

396.9KB

SUBMISSION DATE

Oct 26, 2023 4:37 PM GMT+5:30

REPORT DATE

Oct 26, 2023 4:37 PM GMT+5:30

20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 12% Internet database
- Crossref database
- 18% Submitted Works database
- 11% Publications database
- Crossref Posted Content database

Enhancing Property Selling with Predictive Analytics using Random Forest

Anandhu Biju
Master of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, Kerala
anandhubiju.official@gmail.com

Sruthimol Kurian
Master of Computer Application
Amal Jyothi College of Engineering
Kanjirappally, Kerala
sruthimolkurian@amaljyothi.ac.in

***Abstract*— "Enhancing the Process of Property Selling with Predictive Analytics is a powerful tool that empowers property owners to make well-informed decisions when selling their properties, a crucial aspect of property management and the real estate business. This proposed model leverages predictive analytics to optimize the process of selling real estate. Its primary objective is to provide property owners with an estimate of the time required to close a property deal. The model utilizes advanced regression techniques, particularly Random Forest Regressors, to determine the optimal timing for property sales. It takes into consideration various customer preferences and property characteristics, including property type, the number of rooms and bathrooms, total room count, floor level, price range, property age, and additional features like furniture, refrigeration, and renovation status.**

Furthermore, it incorporates property-related data such as recent renovations, access to water sources, proximity to major roads and grocery stores, all categorized by region ratings (ranging from 0-1, 2-5, 5-10). In essence, this model streamlines the decision-making process for property owners, saving them time and increasing the overall efficiency of real estate transactions.".

Keywords— Random Forest, Machine Learning.

I. INTRODUCTION

In the ever-changing property management and real estate industry, the importance of informed decision making cannot be overstated.

This article presents a comprehensive model that revolutionizes the real estate sales process by seamlessly integrating predictive analytics. The main goal of this model is to provide homeowners with an important estimate of the time it will take to sell their property. Additionally, it offers personalized recommendations to improve the market value of the property. At its core, this innovative solution uses random forest regression tools, a sophisticated machine learning technique, to determine the optimal time to bring assets to market. User-defined options include property type, number of bedrooms and bathrooms, total number of rooms, floors, price range, age of ownership and property features such as furniture, air conditioning and Available parking spaces are taken into account. However, the scope of the model goes beyond these parameters, including essential information about the property such as recent renovations, available water supply and distance from main roads and supermarkets , classified into regional scores from 0 to 10. This holistic approach ensures a comprehensive approach and differentiated real estate valuation, providing owners with the necessary information about market dynamics. It is important to note that the model is not limited to making predictions .

In summary, this model makes significant contributions to the real estate industry. It simplifies and streamlines decision-making for property owners, providing critical insights to make informed decisions. This article delves deep into the intricacies of the model, offering a comprehensive understanding of its components, formatting, styling, and key keyword insertion, serving as an innovative beacon in the dynamic real estate industry.

II. RELATED WORKS

Sumanth Mysore and team [1] used diverse machine learning algorithms, including Support Vector Regression and CatBoost Regressor, to predict house prices. They considered factors like room count, square footage, and proximity to employment centers. The choice of regression model significantly affected prediction accuracy

Anurag Sinha [2] used a set of machine learning algorithms to estimate real estate prices. Ordinary least Squares techniques are used in this research. The square footage of each property played an important role in determining the price, as well as the size of the lot, the number of rooms, and the number of bathrooms in the home. Several feature extraction metrics are used to predict house prices in this work, these variables are referred to as feature datasets. According to the results, it had a significant impact on the placement of the dwelling.

Nguyen and Shahabi [3] used a data-driven approach to predict real estate prices using machine learning. Their study focused on using different machine learning algorithms to estimate real estate prices. They mainly used Ordinary Least Squares methods. Key factors in determining property prices included the square footage of each property, lot size, number of rooms, and number of bathrooms in the homes. The research also included the extraction of house price prediction features known as feature datasets, and the results showed that these variables significantly influenced property valuation

Abhijit Sarma et al. [4] presented a study titled "Using Machine Learning and Neural Networks to Predict House Prices." The study proposes a method to accurately estimate the value of real estate properties. The system incorporates networks, linear regression, forest regression. Boosted regression techniques. This approach is appealing to customers as it provides results and minimizes the risk of making an investment, in a house.

Sifei Lu [5] proposed the use of a "hybrid regression method" for asset valuation. This study uses limited data features to evaluate the improvement feature design approach. The recent "Kaggle" Challenge "House Price: Advanced Regression Techniques" used the proposed method as its central corner. The article aims to measure the fair value of consumers based on their objectives and financial situation

III. Methodology

A. Data Loading and Preprocessing:

The methodology starts with data loading and preprocessing. The code starts by importing the libraries needed for data manipulation and machine learning, including Pandas for data processing and Sklearn for various machine learning modules. Use the pd.read_csv function to retrieve the dataset from her CSV file named main.csv. Categorical features are encoded using a label encoder, paying special attention to columns such as "property_type", "furnished", "air_conditioner", "parking", and "water_available". These transformations are important for preparing data for modeling. Next, the dataset is divided into two segments.

Function scope (X) and target variable (y). The ultimate goal is to predict "sale duration" based on other features in the dataset. The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.

B. Model initialization and training::

The model used in this methodology is a random forest regressor. It is initialized with key parameters including the number of estimators (100) and a random seed for repeatability (42). The model is then trained using the preprocessed data. In this st a random forest regressor is applied to the feature set (X) and the target variable (y). The trained model is important for predicting the sales period of real estate.

C. Model and Label Encoder Saving:

Once the model is trained and tested, it is saved in the file 'seminar.pkl' using the joblib.dump method. This step ensures that the trained models can be used in future applications without having to retrain them. In addition, the label encoders used to preprocess the data are also stored in files, allowing for more consistent data coding in future use cases Each label encoder is stored with a descriptive file name make it easier to check..

D. User Input and Prediction:

The program communicates with the user to estimate how long it will take to sell a property. He gathers detailed information, about the property. This information includes things like the type of property, the value of the number of bedrooms and bathrooms or the parking space available when it was last repaired or how close it is to a road and supermarket well if it has a flow like age and the total number of rooms. All this input, will be formatted in a DataFrame for the user.

E. Predictions and Property Analysis:

The system uses the user's input to make predictions, about how the property will have time to sell. To make these predictions, a model learned from the data is used. Additionally, the law includes an element of analysis of the property itself. The goal is to help real estate owners optimize their sales process.

F. Model Evaluation and Accuracy Metrics:

The next step is to evaluate the accuracy and performance of the model. The mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared score (R2) are calculated. These metrics provide valuable insight into the predictive ability of the model and the quality of the predictions produced. Metrics serve as benchmarks to measure the effectiveness of your model.

G. Model Accuracy Metrics and Property Tips Display:

And then provide a comprehensive analysis of the model's performance and property specific advice for selling properties. Precision indexes, including mean absolute error (MAE), mean square error (MSE), root mean error of square (RMSE), and R-squared (R2) scores, provide valuable insights into how the model its actual data and its forecasts match well accuracy. Additionally, we use these assessments to provide property-specific understanding based on factors such as affluence, parking, water availability, renovation history, etc. They help develop and make decisions.

IV. Build Model

In our efforts to build simple and accurate machine learning models for predicting property sales timing, we carefully follow a variety of well-defined methods to ensure their efficiency. These steps include data preprocessing, feature engineering, model selection and training, rigorous analysis, and fine-tuning. With this planning approach, we aim to provide you with a valuable tool that helps you make informed real estate decisions.

A. Importing Libraries

```
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import joblib
import os
import matplotlib.pyplot as plt
```

a) pandas:

A Python data manipulation and analysis library that provides data structures and functions for structured data processing.

b) RandomForestRegressor from sklearn.ensemble:

Machine learning algorithms used for regression tasks that leverage a collection of decision trees to predict numbers.

c) LabelEncoder from sklearn.preprocessing:

A tool that encodes categorical features into numbers that make them suitable for machine learning models.

d) mean_absolute_error,mean_squared_error, r2_score from sklearn.metrics:

Functions for assessing regression model performance, including metrics like mean absolute error, mean squared error, and R-squared score.

e) joblib:

A Python library tailored for lightweight pipelining, specifically designed for parallel processing and efficient data sharing

B. *Building the model*

1) Loading the Dataset:

The code begins by loading a dataset from a CSV file named 'main.csv' using the Pandas library. This dataset typically contains information relevant to the property sale duration prediction task.

2) Encoding Categorical Features:

Categorical features in the dataset, such as 'property_type,' 'furnished,' 'air_conditioner,' 'parking,' and 'water_available,' are encoded using LabelEncoder, allowing the model to process them effectively.

```
# Encode categorical features
label_encoders = {}
categorical_columns = ["property_type", "furnished", "air_conditioner", "parking", "water_available"]
for col in categorical_columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le
```

3) Data Preparation:

The dataset is prepared for model training by creating two datasets, 'X' and 'y.' 'X' contains all the columns from the original dataset except 'sale_duration,' as it serves as the target variable. 'y' contains the sale duration values.

```
# Split the data into features (X) and the target variable (y)
X = data.drop('sale_duration', axis=1)
y = data['sale_duration']
```

4) Data Splitting:

To evaluate the model's performance, the dataset is further divided into two subsets: the training set and the testing set. Approximately 80% of the data is allocated for the training set, while the remaining 20% is reserved for testing. This split allows for assessing how well the model generalizes to new, unseen data. The 'random state' parameter is set to ensure reproducibility of the data split.

5) Model Training:

The Random Forest Regressor model is trained on the training data (X_train and y_train) using the fit method.

```
# Initialize and train the Random Forest Regressor on the training set
rf_regressor = RandomForestRegressor(n_estimators=100, random_state=42)
rf_regressor.fit(X_train, y_train)
```

6) Model Serialization:

To ensure the model's persistence and reusability, the trained Random Forest Regressor model is serialized and saved to a PKL (joblib) file located in the "models" directory with the filename 'seminar.pkl.' This enables easy loading and reutilization of the model for making predictions without the need for retraining. Additionally, label encoders for categorical features are saved with appropriate filenames, and all label encoders are collectively stored in a single PKL file, 'label_encoders.pkl,' for convenient access in future applications.

C. *User Input and Data Preparation:*

This phase initiates by gathering crucial property attributes through user input. These attributes encompass property type, bedroom and bathroom counts, price, and binary options regarding furnishings, air conditioning, and parking availability. Moreover, details about the last renovation in a specified range and water availability are recorded. Numeric data such as distances, property age, room count, and floor level are also collected. The input is structured into a DataFrame, aligning it with the model's requirements, paving the way for precise property sale duration predictions.

```
# Define a function to get user input and predict sale_duration
def predict_sale_duration():
    print("Predict Sale Duration for a Property:")
    property_type = input("Property Type (e.g., Apartment, House, Villa, Commercial, Garage): ")
    num_bedrooms = int(input("Number of Bedrooms: "))
    num_bathrooms = int(input("Number of Bathrooms: "))
    price = int(input("Price: "))
    furnished = input("Furnished (Yes/No): ").strip().lower()
    furnished = 1 if furnished == 'yes' else 0
    air_conditioner = input("Air Conditioner (Yes/No): ").strip().lower()
    air_conditioner = 1 if air_conditioner == 'yes' else 0
    parking = input("Parking (Yes/No): ").strip().lower()
    parking = 1 if parking == 'yes' else 0
    last_renovation_years_ago_input = input("Last Renovation (Years Ago, e.g., 2-5): ")
    water_available = input("Water Available (Yes/No): ").strip().lower()
    distance_to_major_road_input = input("Distance to Major Road (e.g., 1-5 km): ")
    distance_to_supermarket_input = input("Distance to Supermarket (e.g., 1-5 km): ")
    property_age_input = input("Property Age (in Years, e.g., 6-10): ")
    total_rooms = int(input("Total Rooms: "))
    # Create a DataFrame with user input
    user_input = pd.DataFrame({
        'property_type': [property_type],
        'num_bedrooms': [num_bedrooms],
        'num_bathrooms': [num_bathrooms],
        'price': [price],
        'furnished': [furnished],
        'air_conditioner': [air_conditioner],
        'parking': [parking],
        'last_renovation_years_ago': [last_renovation_years_ago],
        'water_available': [water_available],
        'distance_to_major_road': [distance_to_major_road],
        'distance_to_supermarket': [distance_to_supermarket],
        'property_age': [property_age],
        'total_rooms': [total_rooms],
        'floor': [floor],
    })
```

D. *Predictions and Model Evaluation:*

In this section, the capabilities of the prediction model are presented. The model is used to predict the sale time of an asset based on information provided by the user, thereby allowing evaluation of the model's performance. Key metrics, including Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and R-squared (R2) score, are meticulously calculated. These metrics provide invaluable insight into the model's accuracy and effectiveness in predicting real estate sale times. In addition, attractive visual representations are created, depicting error distribution through meticulously constructed histograms using the create_error_histogram function. Additionally, a scatter plot depicting actual and predicted sales duration is elegantly presented using the create_scatter_plot function. These visual aids enrich the understanding of model performance. Additionally,

trained label and model encoders are efficiently serialized and stored in the "model" directory, ensuring convenient access to prediction capabilities.

E. Display Results:-

In this section, the model's predictions on asset sale time will be revealed, leveraging user-supplied asset attributes. Model prediction accuracy is widely evaluated using a set of important metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R-squared score (R2). These metrics act as a compass, guiding our understanding of the model's performance and its ability to predict real estate sales times. Additionally, attractive visual representations are designed to enhance understanding of model performance. The carefully constructed histogram with the create_error_histogram function provides a clear picture of the error distribution. Additionally, the scatter plot is presented beautifully using the create_scatter_plot function, illustrating the relationship between actual and forecast sales duration. These visual aids provide a comprehensive view of the model's accuracy and effectiveness in predicting real estate sale times.

```
Model Accuracy Metrics on Test Set:
Mean Absolute Error (MAE): 6.727231920199502
Mean Squared Error (MSE): 93.56638653366582
Root Mean Squared Error (RMSE): 9.672971959727052
R-squared (R2) Score: 0.9095529915573666
```

Fig. 1. Histogram

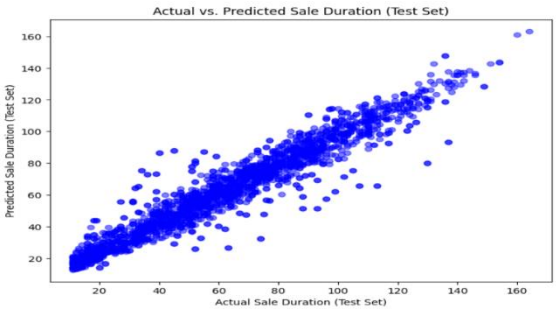

Fig. 2. scatter_plot

V. Application

In this part of the app, users receive personalized recommendations that have been successfully implemented in our project to optimize their real estate listings for effective sales. After receiving the property data provided by the user, the model integrated in our app performs predictive analysis to estimate when the property will be sold. These predictions, generated as part of the project, are combined with actionable insights specific to the user's property. The recommendations we implement cover several aspects, including property type, interior finishes, air conditioning, parking, and water availability. In addition, the project includes renovation suggestions based on recent renovations that are seamlessly integrated into an app. This information, a core component of the project, helps users improve the desirability and marketability of their properties, enabling more effective real estate sales. By integrating advanced predictive modeling with user input, the project provides a user-friendly experience that helps real estate sellers make informed decisions and realize the full potential of their property listings.

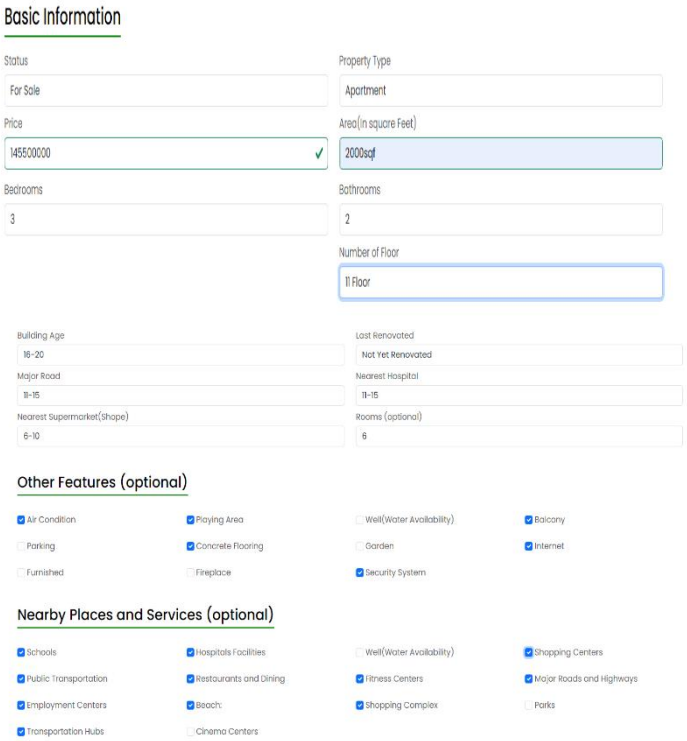

Fig. 1. User input of their scores.

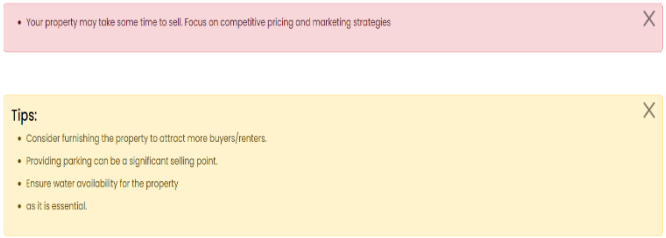

Fig. 2. Output result for the user input

VI. Conclusion

In conclusion, the random forest regressor model described here provides a robust and reliable solution for predicting the sales duration of real estate. By using a dataset enriched with various property attributes, this model exploits the predictive power of ensemble learning. Its strength lies in providing accurate estimates of the time it takes to sell a property.

When we evaluate the performance of this model using metrics such as mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE), and R-squared (R2), it becomes clear that this model provides very accurate predictions. In addition, this model provides real estate sellers and real estate professionals with valuable insights into the marketability of their properties. By automating the prediction of sales duration, it mitigates uncertainty in the real estate market and reduces the need for costly consulting services. It promotes a data-driven and unbiased approach to evaluating the likelihood of property sales, improving decision-making for real estate industry stakeholders.

In short, this random forest regressor model is an important tool for potential real estate sellers to help them gain a deeper understanding of their property's market dynamics and facilitate informed real estate decisions.

VII. References

[1] Mysore, Abhinay Muthineni, Vaishnavi Nandikandi, Sudersan Behera. "Prediction of House Prices Using Machine Learning." International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 10, Issue VI, June 2022

[2] Anand G. Rawooli, Dattatray V. Rogye, Sainath G. Rane, and Dr. Vinayak A. Bharad. "House Price Prediction using a Machine Learning Model: A Survey of Literature." IRE Journal (May 2021).

[3] Nguyen, H. T., & Shahabi, C. (2016). "A data-driven approach for predicting housing prices using machine learning techniques."

[4] Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair - "Housing Price Prediction Using Machine Learning and Neural Networks" 2018, IEEE.

[5] Sifei Lu, Zengxiang Li, Zheng Qin , Xulei Yang , Rick Siow Mong Goh - "A hybrid regression technique for house prices prediction" 2017,IEEE.

[6] Adeyemo, A. O., Adewumi, A. O., & Ayo, C. K. (2015). A comparison of regression models for prediction of house prices. Journal of Real Estate Literature, 23(1), 63-83.

[7]Anurag Sinha "Utilization Of Machine Learning Models In Real Estate House Price Prediction".[July 2020]

[8] Ding, M., & Liu, Z. (2019). "Housing price prediction using machine learning."

[9] Panagiotidis, T., & Pr[illegible]s, P. (2017). "Forecasting house prices using dynamic model averaging." Journal of Real Estate Finance and Economics, 54(1), 95-112.

[10] Verma, G. K., & Jha, N. N. (2013). Housing price prediction: Parametric and non-parametric approaches. International Journal of Computer Applications, 62(17), 12-18.

[11] Adeyemo, A. O., Adewumi, A. O., & Ayo, C. K. (2015). A comparison of regression models for prediction of house prices. Journal of Real Estate Literature, 23(1), 63-83.

20% Overall Similarity

Top sources found in the following databases:

- 12% Internet database
- Crossref database
- 18% Submitted Works database
- 11% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**Baljap Singh, Jaspreet Singh, Anubhav Kumar, Shikha Gupta. "Investiga...** Crossref	1%
2	**University of Northampton on 2023-09-24** Submitted works	<1%
3	**ijarsct.co.in** Internet	<1%
4	**pergamos.lib.uoa.gr** Internet	<1%
5	**University at Buffalo on 2023-10-13** Submitted works	<1%
6	**Süreyya Özöğür Akyüz, Birsen Eygi Erdogan, Özlem Yıldız, Pınar Karada...** Crossref	<1%
7	**University of Portsmouth on 2023-09-20** Submitted works	<1%
8	**repository.umy.ac.id** Internet	<1%

9	University of Bradford on 2022-11-25 Submitted works	<1%
10	medium.com Internet	<1%
11	irejournals.com Internet	<1%
12	ymerdigital.com Internet	<1%
13	University of Liverpool on 2023-09-14 Submitted works	<1%
14	Madeline Lee, Yee Ser, Ganeshsree Selvachandran, Pham Thong, Le C... Crossref	<1%
15	Texas A & M University, Kingville on 2023-04-19 Submitted works	<1%
16	ijraset.com Internet	<1%
17	University of Bradford on 2023-05-21 Submitted works	<1%
18	ndwrcdp.werf.org Internet	<1%
19	University of Sunderland on 2023-10-13 Submitted works	<1%
20	Morton High School on 2023-10-04 Submitted works	<1%

21	Unizin, LLC on 2021-11-02 Submitted works	<1%
22	researchgate.net Internet	<1%
23	University of Bradford on 2023-09-06 Submitted works	<1%
24	Massey University on 2023-10-16 Submitted works	<1%
25	University of North Texas on 2023-10-14 Submitted works	<1%
26	dev.to Internet	<1%
27	srmist on 2023-09-28 Submitted works	<1%
28	Eastern Gateway Community College on 2023-10-08 Submitted works	<1%
29	Colorado State University Fort Collins on 2021-11-13 Submitted works	<1%
30	MCAST on 2015-05-07 Submitted works	<1%
31	Queensland University of Technology on 2023-09-15 Submitted works	<1%
32	Geetha D. Devenagavi, Bukkapatnam Mohammed Aleem, C Saiswaroo... Crossref	<1%

33	Harrisburg University of Science and Technology on 2019-07-16 Submitted works	<1%
34	London Metropolitan University on 2023-05-12 Submitted works	<1%
35	University of Bradford on 2023-03-27 Submitted works	<1%
36	Whitecliffe College of Art & Design on 2023-08-31 Submitted works	<1%
37	Xiaojie Xu, Yun Zhang. "Office property price index forecasting using n... Crossref	<1%
38	Yefu Chen, Junfeng Jiao, Arya Farahi. "Disparities in affecting factors o... Crossref	<1%
39	ijcseonline.org Internet	<1%
40	frontiersin.org Internet	<1%
41	ijml.org Internet	<1%
42	irjmets.com Internet	<1%
43	National College of Ireland on 2023-08-29 Submitted works	<1%
44	Ton Duc Thang University Publication	<1%

45 University of Wales Institute, Cardiff on 2023-10-18

Submitted works

<1%