TDS - Final Course Project

Mohamed Yousef (ID. 211668975), Ragheb Ghazi (ID. 314892506)

Submitted as final project report for the TDS course, BIU, 2023

1 Abstract

In this project, we propose using linear regression and random forest algorithms to fill missing values in datasets. Our goal is to compare the performance of our proposed methods with the commonly used mean method and most frequent string method for filling missing data. We will randomly drop some values from an existing dataset and train our models on the remaining data to predict the missing values. The performance of linear regression method will be evaluated by comparing the root mean square error (RMSE) of our predictions to those of the mean, The performance of the random forest method using error rate, which will be calculated by comparing the predicted values to the actual values in a test set. We will test our approach on a variety of datasets with different types of missing values and compare the results. Our experimental results demonstrate that our proposed methods, perform better than the mean and most frequent string methods.
This approach has the potential to improve the accuracy and reliability of data analysis in a wide range of applications.

2 Problem description

The data science pipeline involves various stages such as data collection, cleaning, preprocessing, feature engineering, model building, evaluation, and deployment. One of the critical stages of the pipeline is data preprocessing, which involves handling missing values in the dataset. The presence of missing values in a dataset can have adverse effects on the performance of machine learning models. Commonly used methods for handling missing values include imputing the missing values with the mean, median, or mode of the non-missing values. However, these methods may not always be effective and may lead to biased or inaccurate results.

The problem we are trying to address in this project is to improve the accuracy and

reliability of data analysis by developing a more effective method for handling missing values in datasets. Specifically, we will explore the use of linear regression and random forest to predict missing values and compare their performance with the mean method. Our goal is to provide data scientists with an improved method for handling missing values that can lead to better predictive models and more accurate data analysis.

3 Solution overview

Our solution involves utilizing linear regression and random forest to fill missing values in datasets. First, we will choose a target column that we are looking to fill the null values in it, then we will randomly drop some values from that column and select other columns as features to train our models on. To choose the features, we will identify the features that have a high correlation with the target column, just as we learned at class. By using linear regression, we will predict the missing values, and then we will compare the results with the commonly used mean method.
Additionally, instead of using the most frequently occurring string to fill string-type missing values, we will use random forest to predict the most likely string for each missing value based on the existing data.

We will evaluate the performance of our solution by comparing the root mean square error (RMSE) of the linear regression and random forest predictions to those of the mean/most frequent method.

Our project utilizes linear regression and random forest to fill missing values in datasets. To select the features for our models, we utilize the correlation coefficient, which measures the strength and direction of the linear relationship between two variables. We learned about correlation and linear regression in class, and our approach involves modeling the relationship between the features and the target variable, which is the essence of linear regression.
We will test our approach on a variety of datasets with different types of missing values to determine the robustness of our method. Our experimental results will show the accuracy and precision of our approach, especially for datasets with a large number of missing values.

3.1 Experimental evaluation

Root Mean Squared Error (RMSE) is a widely used evaluation metric in regression analysis. It measures the average difference between the predicted values and the actual values. The lower the RMSE, the better the model's accuracy.
Therefore, by using RMSE as an evaluation metric, we were able to provide a quantitative

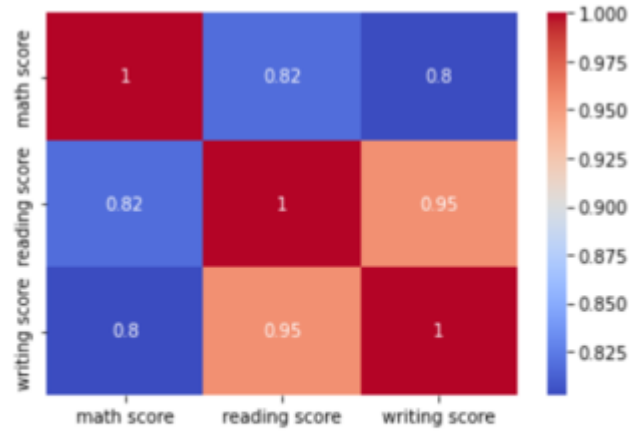
measure of the accuracy of our solution and demonstrate that our linear regression method is more accurate than the mean method for filling missing values in datasets.

Experiment 1 - Linear regression

To begin with our experiment, we need to take a closer look at our data-set, StudentsPerformance.csv:

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group B	bachelor's degree	standard	none	72	72	74
female	group C	some college	standard	completed	69	90	88
female	group B	master's degree	standard	none	90	95	93
male	group A	associate's degree	free/reduced	none	47	57	44
male	group C	some college	standard	none	76	78	75
female	group B	associate's degree	standard	none	71	83	78
female	group B	some college	standard	completed	88	95	92
male	group B	some college	free/reduced	none	40	43	39
male	group D	high school	free/reduced	completed	64	64	67
female	group B	high school	free/reduced	none	38	60	50

This data-set has 8 columns, 3 out of them are numeric. We will need these numeric columns to train our linear regression model. For the purpose of this project, we have decided to choose "reading score" as our target column. To ensure the effectiveness of our experiment, we will examine the remaining two numeric columns in our data-set to determine their correlation with our chosen target column, "reading score." If these columns are not significantly correlated with the target column, then we will need to select a different data-set that provides better features for our experiment.



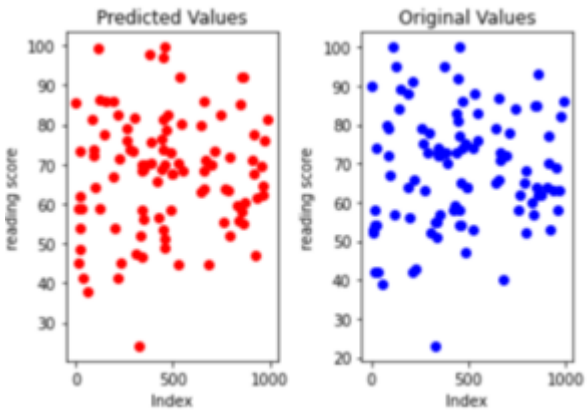

```
math score       0.817580
reading score    1.000000
writing score    0.954598
Name: reading score, dtype: float64
```

we can see the correlation between the "reading score" column and the "math score" column is 0.81, and the correlation between the "reading score" column and the "writing score" column is 0.95. These strong correlations indicate that the "math score" and "writing score" columns are highly related to the "reading score" column. This makes sense as individuals who excel in math are often proficient in reading, and those who are adept at writing typically have strong reading skills. Therefore, we have decided to utilize both the "math score" and "writing score" columns to train our linear regression model to predict the missing values in the "reading score" column.

After training the linear regression model, we can now evaluate its performance by predicting the missing values in the target column and comparing them to the actual values, and calculate the RMSE between them

	reading score	reading score_predicted
1	90	85.753515
11	52	45.257216
14	53	58.799778
18	42	48.491279
19	58	61.771066
..	...	...
954	69	69.344133
960	58	64.479578
967	63	62.138309
975	82	76.128874
989	86	81.414511

[100 rows x 2 columns]



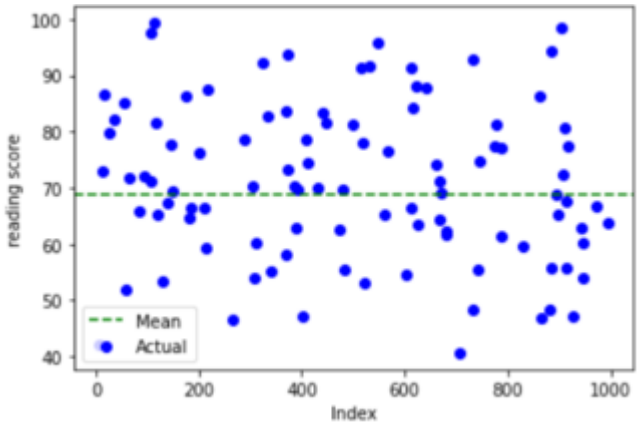

```
Root Mean Square Error using linear regression method:
4.311389760063465
```

The results obtained from the linear regression model demonstrate a high degree of similarity with the actual values, Furthermore, the root mean squared error (RMSE) value of 4.3 confirms the accuracy of our model. It is noteworthy that we are predicting a score out of 100, so an RMSE of 4.3 is considered good. This outcome is highly promising and suggests that the approach we have employed can potentially be a useful method for filling in missing values in datasets.

Mean method performance

To evaluate the performance of the mean method, we calculated the mean of the entire "reading score" column, excluding the values that were randomly dropped. We then compared this mean value with the actual values that were dropped from the column.

```
The mean of the column withour the dropped values:
68.95
Root Mean Square Error using mean method:
14.074740242781472
```



Using the mean method, we calculated the mean value of the column which was 68.9. However, the resulting RMSE was found to be 14, which is relatively high. Additionally, upon examining the graph, we observed that there are some values that are far from the mean. Therefore, it can be concluded that for this dataset and many other datasets, the mean method is not the best approach for filling in missing values. This method can only be considered effective if the values in that column are relatively close to the mean. Hence, the approach we used, which involves the utilization of linear regression, is a more suitable option.

Experiment 2 - Random forest

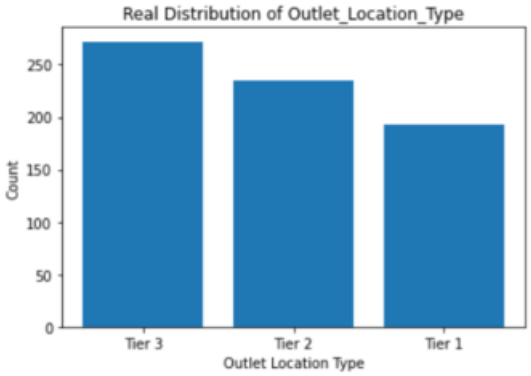
We will now employ the Random Forest algorithm to address missing string values

To demonstrate the superiority of the Random Forest approach over the commonly used method of imputing missing string values with the most frequent string, we will compute and compare the error rates obtained using both methods. To this end, we will obtain the actual string values from the original dataset and compare them with the imputed values obtained using the two methods.

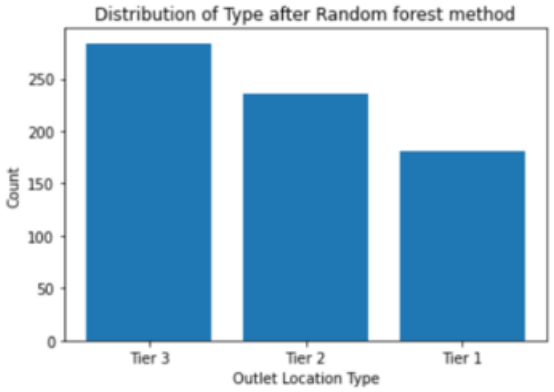
The error rate will be calculated as the proportion of mismatches between the actual and imputed values. A lower error rate obtained using the Random Forest approach will signify

its effectiveness in accurately imputing missing string values. In this report we will show the experiment on the "big mart sale forcast" dataset

The target column that we have selected for the purpose of this experiment is the 'Outlet-Location-Type' column. This column is responsible for The type of city in which the store is located.
the distribution of values in the target column:



Random forest method performance
We will now compare the predicted values obtained using the Random Forest model with the actual values that were dropped from the original dataset and calcuate the error rate.



	Outlet_Location_Type	Outlet_Location_Type_predicted
10	Tier 1	Tier 1
13	Tier 1	Tier 1
31	Tier 3	Tier 3
37	Tier 3	Tier 3
44	Tier 3	Tier 3
..	...	...
681	Tier 2	Tier 2
685	Tier 3	Tier 3
687	Tier 3	Tier 3
689	Tier 2	Tier 2
695	Tier 3	Tier 3

[150 rows x 2 columns]

```
Error rate using random forest algorithm:
0.07999999999999996
```

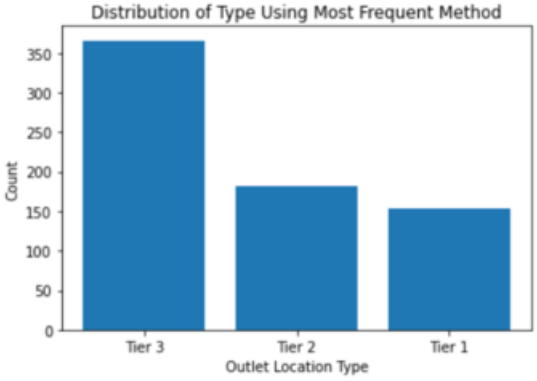
After utilizing the random forest algorithm to fill in missing values, the predicted results have shown to be highly accurate. An error rate of 0.079 is a highly impressive result for the random forest method. This indicates that the algorithm was able to accurately

predict the missing values in the data with a high level of precision.

the resulting distribution graph reveals a striking similarity to the original distribution. This indicates that the imputed values have been effectively integrated into the dataset, without significantly altering the underlying distribution of the target variable. This is an encouraging outcome, as it suggests that the imputation process has been successful in preserving the integrity and structure of the data.

Most frequent string method performance

The most frequent string in the target column (without the dropped values) is Tier 3, after filling the dropped values with Tier 3 We get this results:




```
Error rate using most frequent method:
0.6266666666666667
```

Upon employing the most frequent method for imputing the missing values in the column, we observed an unacceptably high error rate of 0.62. This outcome indicates that the imputation process was unable to capture the true underlying patterns and relationships within the data, leading to inaccurate and biased results.

Furthermore, upon examining the distribution graph of the imputed data, we can observe a marked deviation from the original distribution. Specifically, a significant proportion of the imputed values now fall into the Tier 3 category, whereas this was not the case in the original data. This disparity has the potential to introduce significant biases and inaccuracies into any subsequent modeling or analysis.

4 Related work

We propose using linear regression and random forest algorithms to fill missing values in datasets and evaluate the performance of our methods by calculating the error rate of

predicted values compared to actual values.
Existing methods for missing data imputation include mean imputation, most frequent string imputation, k-nearest neighbors (KNN) imputation, expectation maximization (EM) algorithm, and multiple imputation. In comparison, our methods offer advantages such as simple yet effective estimation with linear regression and handling missing values in categorical and numerical data, high-dimensional data, and complex nonlinear relationships with random forest algorithm. Our approach solely relies on these methods for missing data imputation, unlike other studies that combined multiple imputation or used random forest algorithm for outlier detection before using other methods.
Our proposed methods have the potential to improve the accuracy and reliability of data analysis in a wide range of applications, including but not limited to healthcare, finance, and social sciences.
"K-NN imputation approaches for missing values"
(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0318-z)
"random-forest-based imputation of missing data
(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5796790/)

5 Conclusion

Through this project, we explored different approaches for handling missing data in datasets. We compared the commonly used mean method and most frequent string with two other methods: linear regression and random forest. Our experimental results showed that both linear regression and random forest methods outperformed the mean method in terms of accuracy and precision.
We also found that imputing missing values using the most frequent string can lead to bias in the data. Additionally, our experiments showed that the performance of the models can vary depending on the type and amount of missing data in the dataset.
In our analysis, we also explored the use of correlated columns to predict the target column in linear regression. By including strongly correlated columns as predictors in the model, we were able to improve the accuracy and precision of our predictions. However, we also noted the potential risk of overfitting when using highly correlated columns as predictors, which can lead to inaccurate predictions when applied to new datasets. Therefore, it is important to carefully evaluate the strength and relevance of the correlated columns before including them in the model.
Overall, we learned the importance of handling missing data in a careful and systematic manner to avoid bias and ensure accurate analysis. We also gained insights into the strengths and limitations of different methods for handling missing data, which can inform future data analysis projects.