|  |  |
| --- | --- |
|  | **Introduction to Business Data Analytics** |

|  |  |
| --- | --- |
| **Homework #4 Part 2** |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(put your names above (incl. any nicknames))

Note: This is a team homework assignment. Discussing this homework with your classmates outside your MSBA team is a **violation** of the Honor Code. If you **borrow code** from somewhere else, please add a comment in your code to **make it clear** what the source of the code is (e.g., a URL would sufficient). If you borrow code and you don’t provide the source, it is a violation of the Honor Code.

Total grade: \_\_\_\_\_\_\_ out of \_\_\_145\_\_\_ points

**(145 points) Use numeric prediction techniques to build a predictive model for the HW4.xlsx dataset. This dataset is provided on Canvas and contains data about whether or not different consumers made a purchase in response to a test mailing of a certain catalog and, in case of a purchase, how much money each consumer spent. The data file has a brief description of all the attributes in a separate worksheet. We would like to build predictive models to predict how much will the customers spend; Spending is the target variable (numeric value: amount spent).**

**Use Python for this exercise.**

**Whenever applicable use random state 42 (10 points).**

1. **(50 points) After exploring the data, build numeric prediction models that predict Spending. Use linear regression, k-NN, and regression tree techniques. Briefly discuss the models you have built. Use cross-validation with 10 folds to estimate the generalization performance. Present the results for each of the three techniques and discuss which one yields the best performance.**

[part a is worth 50 points in total:

15 points for exploring the data (i.e., descriptive statistics including min max mean and stdv, visualizations, target variable distribution)

10 points for correctly building linear regression model provide screenshots and explain what you are doing and the corresponding results

10 points for correctly building k-NN model provide screenshots and explain what you are doing and the corresponding results

10 points for correctly building regression tree model provide screenshots and explain what you are doing and the corresponding results

5 points for discussing which of the three models yields the best performance]

1. **(50 points) Engage in feature engineering (i.e., create new features based on existing features) to optimize the performance of linear regression, k-NN, and regression tree techniques. Present the results for each of the three techniques (choose the best performing model for each technique in case you try multiple models) and discuss which of the three yields the best performance. Use cross-validation with 10 folds to estimate the generalization performance. Discuss whether and why the generalization performance was improved or not.**

[part a is worth 50 points in total:

10 points for correctly building the new linear regression model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

10 points for correctly building the new k-NN model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

10 points for correctly building the new regression tree model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

20 points for discussing if the generalization performance was improved or not for each of the techniques (linear regression, kNN, and regression tree) and justifying why it was improved or alternatively why it was not improved]

**Based on the results you've provided, it appears that feature engineering has led to improvements in the generalization performance of the models for Linear Regression, k-NN Regression, and Regression Tree. Let's discuss why these improvements may have occurred:**

**1. Linear Regression:**

**Before Feature Engineering (MAE: 70.50): Linear regression had a relatively high MAE, which suggests that the model was not fitting the data well.**

**After Feature Engineering (MAE: 57.00): After feature engineering, the MAE decreased to 57.00, indicating that the model's ability to predict the target variable (Spending) improved.**

**Possible Justification: The polynomial features created during feature engineering may have introduced nonlinear relationships between the input features and the target variable. This allows linear regression to capture more complex patterns in the data, leading to better predictions.**

**2. k-NN Regression:**

**Before Feature Engineering (MAE: 59.16): k-NN regression had a moderate MAE, indicating reasonable performance but with room for improvement.**

**After Feature Engineering (MAE: 58.85): The MAE slightly decreased to 58.85, suggesting a modest improvement.**

**Possible Justification: Feature engineering might have helped in identifying interactions and higher-order relationships in the data, which k-NN can capture. However, the improvement is not substantial, possibly because k-NN is already a flexible model that adapts to data patterns.**

**3. Regression Tree:**

**Before Feature Engineering (MAE: 66.74): The regression tree had the highest MAE, indicating poor generalization.**

**After Feature Engineering (MAE: 62.82): The MAE decreased to 62.82, showing an improvement but still higher than the other models.**

**Possible Justification: Feature engineering could have made the data more suitable for the regression tree model. By creating new features and interactions, it becomes easier for the tree to partition the data effectively. However, the improvement might be limited due to the inherent limitations of regression trees in capturing complex relationships.**

**In summary, feature engineering seems to have generally improved the generalization performance of the models, particularly in the case of Linear Regression and k-NN Regression. The introduction of polynomial features and interactions likely allowed the models to better capture the underlying patterns in the data. However, the improvement may vary depending on the model's flexibility and how well it can adapt to the new data representations. Further fine-tuning and model selection could potentially lead to even better results.**

1. **(35 points) Engage in parameter tuning to optimize the performance of linear regression, k-NN, and regression tree techniques. Use cross-validations with 10 folds to estimate the generalization performance. Present the results for each of the three techniques and discuss which one yields the best performance.**

[part a is worth 35 points in total:

10 points for correctly optimizing at least two parameters for linear regression model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

10 points for correctly optimizing at least two parameters for linear k-NN model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

10 points for correctly optimizing at least two parameters for linear regression tree model and improving the performance as much as possible provide screenshots and explain what you are doing and the corresponding results

5 points for discussing which of the three models yields the best performance]