CA Report for

Intro to AI & ML

Adam Plesca
3rd Year BSHC Computing Student
National College of Ireland
Dublin, Ireland

Abstract [Extra]

In this report, I will be using Machine Learning Techniques such as Linear & Logical regression along with K-Nearest Neighbours (K-NN) on two different datasets that I have chosen. With the purpose of gaining a better understanding of certain problems I wanted to research into. One focused on airline ticket prices while the other focused on healthcare test results. The data was preprocessed with techniques such as normalization, standardization and class balancing. Evaluation metrics for models were Mean Absolute Error (MAE), Mean Squared Error (MSE), R-squared, accuracy, precision, recall, and F1-score. The models performed quite well but there is always room for future improvement.

Introduction [1]

Machine Learning (ML) has become an extremely popular way to solve various predicative tasks. With valuable insight into new ways of viewing data. This project focuses on two tasks: Regression (predicting airline ticket prices) and Classification (predicting healthcare test results). I chose these tasks to show how versatile ML can be. The goal is to apply Linear regression to regression tasks, and Logistic regression with K-NN for classification tasks to see how well they perform. My report is organized as shown below.

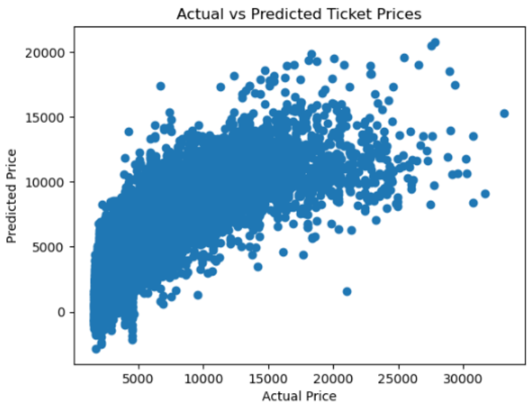
Section Name [Section number]

Motivation [2]

Linear regression has long been used for numerical prediction tasks due to its simplicity and interpretability. K-NN and Logistic regression are popular classification algorithms, used commonly in fields such as healthcare, marketing and so on. Past research has shown that it is very effective. Although there are challenges such as class imbalances and feature scaling which hinder performance. My project seeks to address those challenges using preprocessing and hyperparameter optimization.

Data Statistics [3]

Two datasets were used in my project. Airline dataset contains 12,000 rows of data related to flight prices. Which include information about destination, time, days until departure and so on. The healthcare dataset contains 13,000 rows of health-related test results data from patients. Due to class imbalance the number of non-normal tests were much lower, applying SMOTE helped in increasing accuracy slightly. Both datasets required preprocessing, including handling missing values, scaling features, and balancing classes.



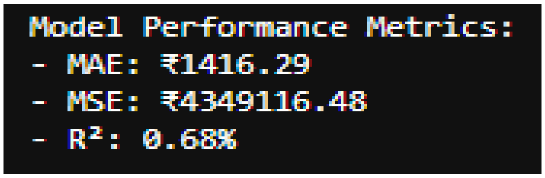
Methodology [4]

Preprocessing involved,
- Inputting missing values using mean or median values.
- Scaling features using standardization (mean = 0, std = 1) and normalization (range [0, 1]).
- Using Synthetic Minority Oversampling Technique (SMOTE) to fix class imbalances.

Algorithms,
1) Linear Regression: Used to predict airline ticket prices, using polynomial features along with interactive terms.
2) Logistic Regression: Used to classify healthcare test results, focusing on accuracy and precision.
3) K-NN: Used to optimize cross-validation for 'k' values and distance metrics (e.g., Euclidean, Manhattan).

Results & Evaluation [5]

Linear Regression Results,

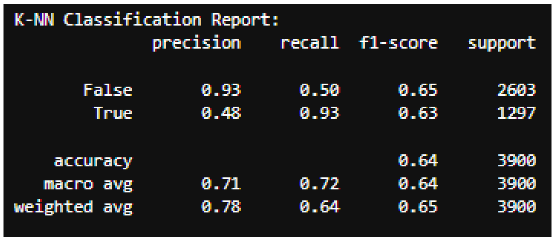


```
Model Performance Metrics:
 - MAE: ₹1416.29
 - MSE: ₹4349116.48
 - R²: 0.68%
```

Mean Absolute Error (MAE) shows that on average the Linear regression model was off by only **₹1416.29** or **€15.91** showing reasonable accuracy**.**
The R² score shows that the model is

accurate by only about roughly 68% which can definitely be improved.

Classification Results,



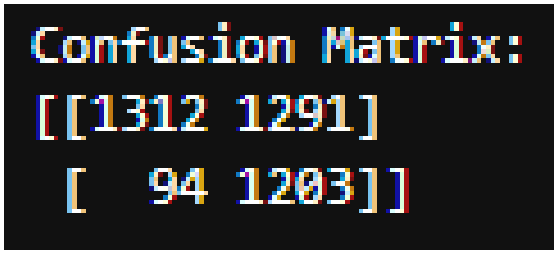
```
K-NN Classification Report:
              precision    recall  f1-score   support

       False       0.93      0.50      0.65      2603
        True       0.48      0.93      0.63      1297

    accuracy                           0.64      3900
   macro avg       0.71      0.72      0.64      3900
weighted avg       0.78      0.64      0.65      3900
```

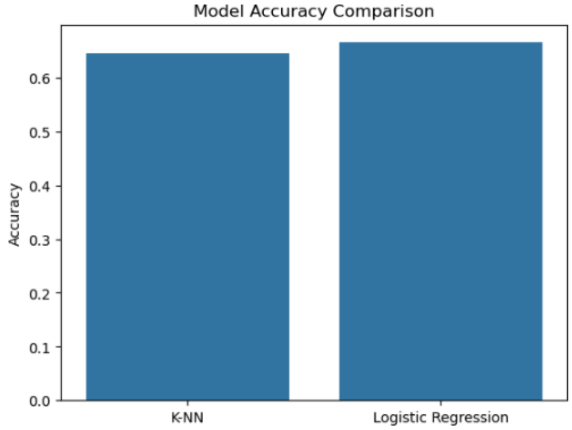
The K-NN model correctly identified 93% of non-normal test results correctly, whereas only 48% of normal ones correct. Which shows high imbalance in prediction accuracy for the K-NN model

Confusion Matrix Results:



```
Confusion Matrix:
[[1312 1291]
 [  94 1203]]
```

Classification Results,



The results show that the Logistical Regression model performs slightly better than the K-NN model in terms of precision and recall, while the Linear Regression model achieves above average prediction accuracy for ticket prices.

Error Analysis [6]

For both Linear regression & K-NN, the models struggled with extremely high values which would be abnormal in most cases. This caused the MAE value to be higher than wanted. Imbalanced classes caused by the healthcare dataset were most likely the cause for the lower score in the K-NN model. Despite using SMOTE. In summary what went wrong reveals that outliers and highly imbalanced classes contributed to errors. Making the performance worse overall. For both the Linear regression model and the K-NN model.

Conclusion [7]

I believe this project showed the application of Linear regression, Logical regression and K-NN for the regression and classification tasks. While the models performed well, there is clear room for improvement. Techniques like ensemble methods, more advanced feature engineering, and hyperparameter tuning can further enhance the results of both the regression and classification task. I believe the key findings were the effectiveness of feature scaling and class balancing in order to improve the model's overall performance.

References [Extra]

Healthcare dataset [Link]

Flight dataset [Link]