Car Price Prediction Model

Mohammad Shahid & Marcelo Enriquez

April 6th, 2023

Professor Sukhjit Singh Sehra

CP322: Machine Learning

Abstract:
The automobile industry is a highly competitive market, and the price of a car is one of the most critical factors in determining the purchase decision of a customer. Accurately predicting the price of a car is essential for car dealerships, manufacturers, and customers. With the vast amount of data available on cars, building a price prediction model using regression analysis can be highly beneficial for its users. The end result of this project is to create a functional and accurate model to give the expected prices of a car given the specifications it has.

The objective of this project is to create a regression model that can predict the price of a car based on specific features such as make, model, year, mileage, fuel type, and transmission type. The model will be trained on a dataset of cars, which includes both new and used cars and the model will also account for inflation over time to provide a more realistic and fitting model.

The methodology for creating the model will involve the following steps:

Data Collection: A dataset of cars has been sourced from third party websites that include the data of popular mid-size sedans from BMW and Toyota ranging from the 20th century to today. This dataset is based on cars sold in Germany. Furthermore, an external dataset named inflation is used to determine the inflation rate on car prices in Europe. By applying the inflation rate to the car prices in our dataset, we can adjust for the impact of inflation and provide a more accurate representation of the real value of the cars at the time of sale. By accounting for inflation, we can make more informed comparisons and analysis of the prices of cars over time.

Data Cleaning and Preprocessing: The dataset will be cleaned and preprocessed to remove any missing or duplicate values, the data will be examined to see if there are any strings in columns where it should only be numbers as well as remove the money symbol to be able to work with the data. Additionally, data given in a range will be updated to display the mean of the range allowing us to work with the data.

Feature Selection: The most relevant features will be selected given the results when various techniques such as correlation analysis with respect to the target variable, the analysis of pair plots as well as the feature importance function from sklearn on a random forest regressor model to obtain the top 10 features. Furthermore, Random Forest regressor determines feature importance, sorts the features by importance, and then selects the top k features using the SelectKBest function and f_regression score function. In this case, we will choose 4 features that are good out of the result we got from the 10 features.

Model Selection: Various regression models such as Linear Regression and Random Forest Regression will be evaluated to determine the best model that provides the highest accuracy. additionally, testing with hyperparameters and ridge regularization will be completed, and the highest r squared score will be selected

Model Evaluation: The performance of the selected model will be evaluated using various metrics such as R-squared, MAE and RMSE.

The outcome of this project will be a robust and accurate car price prediction model that can be used by car dealerships, manufacturers, and customers to estimate the price of a car accurately. The model will provide insights into the factors that influence car prices, and it will assist customers in making informed purchase decisions.

Table of Contents:

Introduction:

The automotive industry is one of the largest and most dynamic industries in the world, with constant changes in technology, consumer preferences, and economic conditions. One critical aspect of the automotive industry is pricing, which can significantly impact the success of car manufacturers, dealerships, and consumers. Accurately predicting car prices is essential for pricing strategies, inventory management, and consumer decision-making.

Regression models are a class of machine learning (ML) models used to predict continuous numerical values. In this project, we will focus on the linear and random forrest regression models, widely used and simple regression models. The linear regression model is appropriate for predicting car prices because it can capture the linear relationship between independent variables (e.g., make, model, year, mileage, and condition) and the dependent variable (price). The aim of this document is to present the process of building a regression ML model that predicts the price of cars. We will explore the various steps involved in building the model, including data collection, data cleaning, feature engineering, and model selection.

We obtained the data used for this project from a publicly available data set containing information on various car models. The dataset includes features such as make, model, year, mileage, and Used price, as well as the target variable (price). We explain how we processed the data and selected relevant features for the model. In order to add complexity and external factors to our model, we included the most recent dataset containing the Harmonized Index of Consumer Prices (HICP). Once the data is collected, we explain how we cleaned the data and prepared it for analysis.

Feature engineering involves creating new features from the existing features that may be more informative and relevant for predicting the target variable. We explain how we selected the best features and created new ones in order to increase the model's predictive power.

We discuss the model selection process and the hyperparameter tuning, which involves selecting the optimal values for the model's parameters to maximize its performance. We also explain the model evaluation metrics used to assess the model's performance, such as mean absolute error, and R-squared.

Once selected, we present the model's performance on a test dataset and compare it to other models. We discuss the strengths and weaknesses of the model and identify potential areas for improvement.

In conclusion, this project demonstrated the process of building a regression ML model that predicts the price of cars. The linear regression model was appropriate for predicting car prices, as it can capture the linear relationship between independent variables and the target variable. The project highlights the importance of accurately predicting car prices in the automotive industry and provides insights into its potential applications. Future work may involve exploring other regression models or incorporating additional features to improve the model's predictive power.

Background:

Machine learning (ML) has revolutionized the field of data science and has become an essential tool for predicting and analyzing data. Regression models are a class of ML models that are used to predict continuous numerical values; that is, a value that has infinite possibilities. The main goal of regression is to find the relationship between the independent variables and the target variable, allowing us to predict the target variable's value for new data.

Types of Regression Models: There are several types of regression models, each with its strengths and weaknesses. Some of the most common types of regression models include linear regression, logistic regression, polynomial regression, ridge regression, lasso regression, and elastic net regression. Within this project, we chose to use linear, ridge and random forest regression models to analyze the dataset.

Linear Regression Model: Linear regression is a straightforward regression model that assumes a linear relationship between the independent variables and the target variable. The model tries to find the best line that fits the data by minimizing the sum of squared errors between the predicted values and the real values. The line's equation is of the form $y = mx + b$, where y is the target variable, x is the independent variable, m is the slope, and b is the intercept. This model can be used for various purposes, such as prediction, estimation, and hypothesis testing. It is widely used in fields such as economics, finance, social sciences, and engineering.

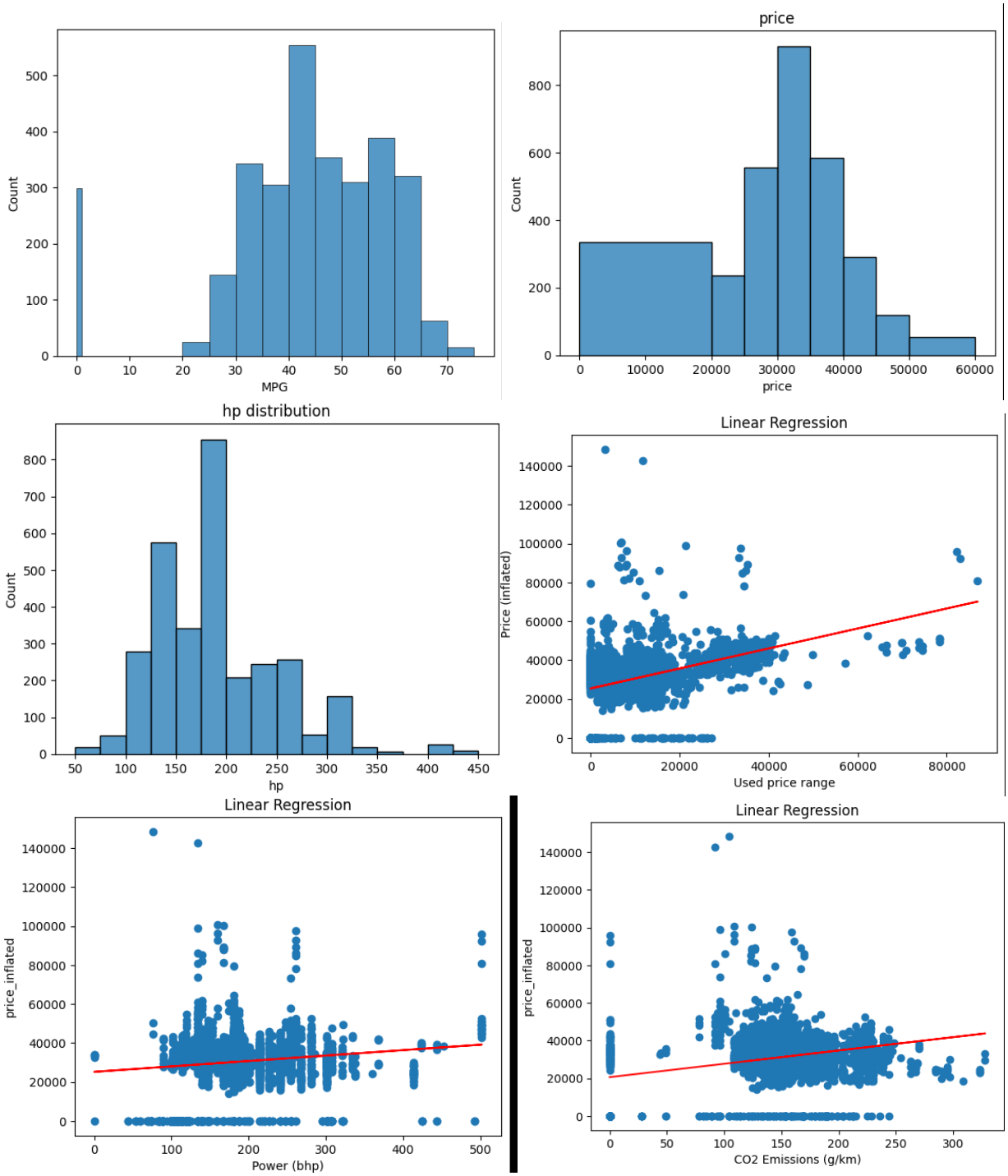
Ridge Regression Model: Ridge Regression is a type of regression that is used when the independent variables in the model are highly correlated with each other. In a regular linear regression model, when two or more independent variables are highly correlated, it becomes difficult to determine the contribution of each individual variable to the dependent variable. This is where Ridge Regression comes into the picture. Ridge Regression is particularly useful when dealing with datasets with many independent variables, as it helps to reduce the risk of overfitting, by adding a penalty term to the objective function, Ridge Regression reduces the complexity of the model and prevents overfitting.

Random Forest Model: Random Forest is a popular and powerful machine learning algorithm that falls under the category of ensemble learning, a technique that combines multiple models to produce better results than any individual model. Random forest works by building several decision trees based on subsets of the training data and features. Then, the trees' predictions are combined to create a final prediction. Each decision tree in the random forest is built using a random subset of the features and data. This process is known as bootstrap aggregating, or bagging. The decision trees are less likely to overfit the training data and can generalize better to new data. Random forest can handle both regression and classification problems and is a flexible model that can handle many features and a wide range of data types. The model's outputs are easy to interpret, and it can identify the most important features for prediction Which is important for our use. Additionally, random forest is a robust model that can handle missing data and outliers in the dataset.

Data Collection:
Data collection is the process of gathering relevant data from various sources. The data sources can include surveys, experiments, databases, and public sources. The data set we used is a sample car database that includes information about cars sold in the United Kingdom. The data set includes variables such as make, model, year, horsepower and fuel type. Reviewing the initial dataset, we decided the variability of cars was lacking and wanted to increase the amount of makes and models to not only increase our accuracy within the dataset but also improve the accuracy with cars outside the dataset by giving the training model more diverse data to learn from. Also, in order to make the model more competitive we have included the HICP index in prices, this makes the older MSRP models more realistic after inflation.

Data Cleaning: Data cleaning is an essential step in the data preprocessing stage. The data cleaning process involves identifying and handling missing values, removing outliers, and dealing with inconsistencies in the data. In the case of the car database, the data cleaning process had numerous revisions as the project developed. The start of preprocessing was filling any null values with 0. and removing characters such as "-" from the data and replacing them with 0, This would allow the columns to be used and manipulated. We then had to update the price columns, removing the British pound symbol for new data creation and manipulation. Additionally, certain columns needed to be updated as they were given in a range, our solution was to get the mean of the range replace the range with that updated value. Lastly, Prior to working on the model we removed all non-continuous variables as they are incompatible with our models. This would be things such as make, model, trim and descriptions. With the inclusion of another dataset to have more cars, there were a few conversions that needed to take place. For one, the added data had its power measured in a different unit than the first, we had to convert each into the right unit (bhp). At the same time, the fuel consumption was measured at L/100km whereas we needed it to be in miles per gallon (MPG).

Data visualization:
Before selecting features and running the data through models, we took numerous parameters and looked at their distribution as well as their linear regression graph with respect to the target variable. Doing so gave us knowledge on the relationships and correlation within the dataset prior to selecting features giving us a more educated and informed view in the later stages of development.



Images a – f: distribution and regression of expected important variables given target of price

Feature Engineering: Feature engineering is the process of creating new variables or features from the existing variables in the data set. Feature engineering is important because it can help improve the performance of the machine learning model. In our case, we were initially unhappy with the accuracy of each model given our dataset. As a result, we decided to calculate the rate of inflation as a decimal, given the inflated price and new price given in the data. We named this feature as “inflation rate” and it gave us a much better and accurate models.

Feature Selection: Feature selection is the process of identifying the most relevant variables or features for the machine learning model. Feature selection is important because it helps to reduce the complexity of the model, and it can improve the model's performance by reducing the risk of overfitting. We decided to start with a correlation analysis from all variables to our target variable of “price inflated”. Looking at the matrix we were able to visually determine which variables have low and high correlation to the target.

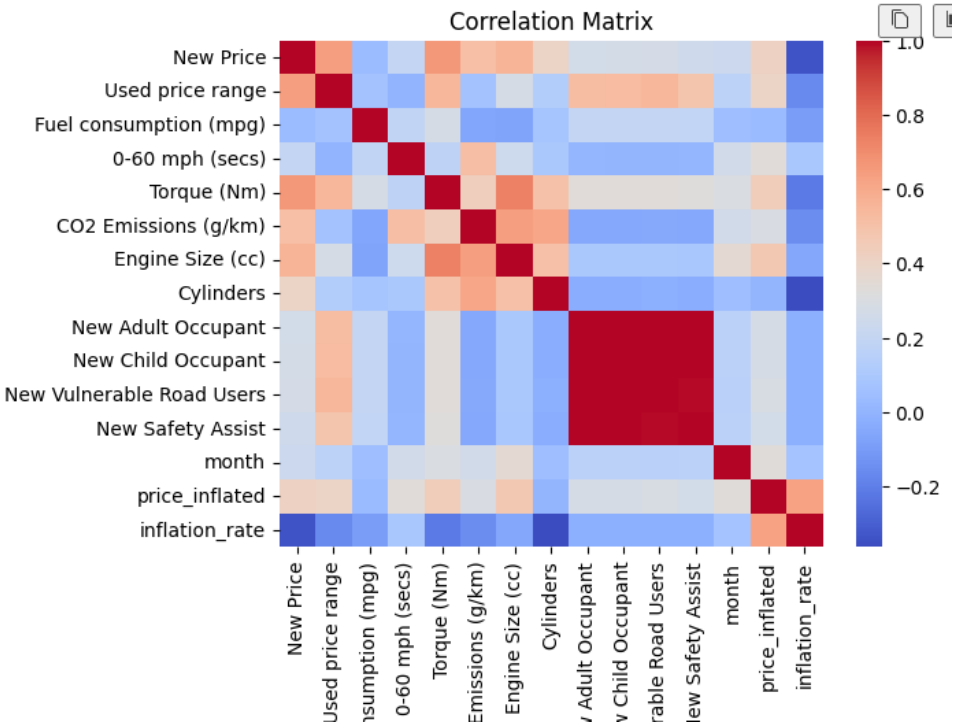


Image g: Correlation matrix

In addition to the correlation analysis, we also used pair pots to visualize the correlation. Pair plots can be used to visualize the relationship between each feature and the target variable. This helps us identify non-linear relationships that we could not see from the correlation matrix alone.

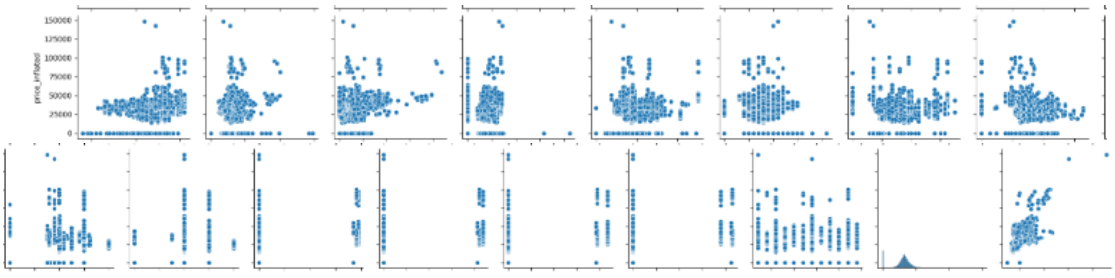


Image h: pair plots where price inflated is the y axis (target variable)

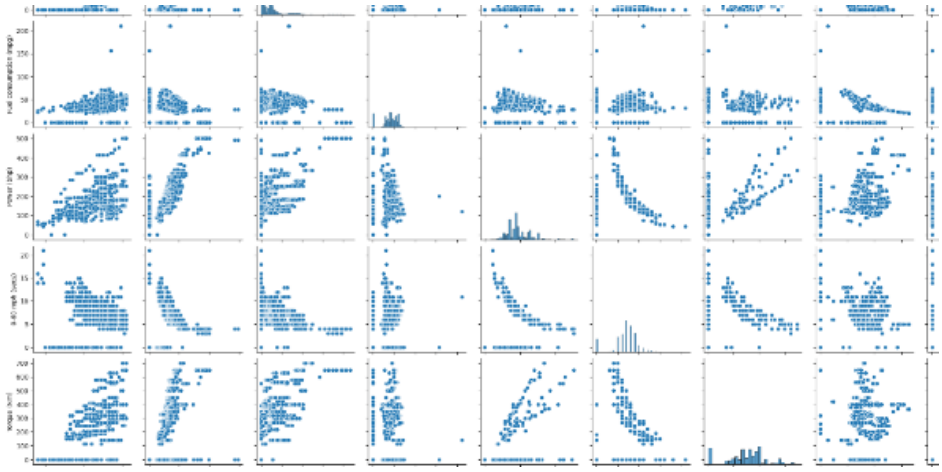


Image I: pair plots of different features showing highly correlated relationships

While it was good to see the correlations, we also wanted to verify them using computational methods. To do this, we fit our data into a Random Forest regressor model, this allowed us to identify k number of features based on importance with respect to our target. After identifying the most relevant features using correlation analysis, pair plots, and random forest regressor, we use the SelectKBest function provided by the scikit-learn library. We used the Random Forrest regressor to identify the top 10 features and then selected the best 4 using the function.Once we have selected the top 4 features using SelectKBest, we then use these features to train a machine learning model, such as a linear regression model, a decision tree, or a random forest. By selecting only the most relevant features, we improve the model's performance and reduce the risk of overfitting.

Train-Test Split: The train-test split is the process of splitting the data set into two subsets - a training set and a testing set. The training set is used to train the machine learning model, and the testing set is used to evaluate the performance of the model. Typically, the training set is a larger subset of the data, and the testing set is a smaller subset. In the case of the car database, we started with a 20% test split but realized that increasing the size of the testing set improved the accuracy of the model. Therefore, we increased the test split to 30% or 40% to obtain a more reliable estimate of the model's performance.

Model Building:
The model selection process is a critical step in machine learning that involves selecting the best model from a set of candidate models, we used linear regression, random forest regression, and ridge regression as candidate models to predict the price inflated variable.

To select the best model, we first trained each model on the training set using scikit-learn's fit function. We then evaluated the performance of each model on the testing set using various metrics, such as mean squared error, mean absolute error, and R-squared.

After evaluating the performance of each model, we found that random forest regression had the highest R-squared accuracy on the testing set, indicating that it performed the best out of the three models. Therefore, we selected random forest regression as the final model for predicting the price inflated variable.

Hyperparameter Tuning & Cross-validation with GridSearch:
Hyperparameter tuning is the process of selecting the best set of hyperparameters for a given machine learning algorithm to optimize its performance on a given dataset. Hyperparameters are the parameters that are not learned by the model during training but are set by the user before training the model. They control the learning process and the complexity of the model, and have a significant impact on the model's performance. In our code, hyperparameter tuning is performed for the Random Forest Regression model using GridSearchCV. GridSearchCV is a technique that exhaustively searches for the best combination of hyperparameters from a given set of hyperparameters by evaluating the performance of the model on a cross-validation set. This process involves partitioning the data into K folds, where K is typically set to 5 or 10. For each fold, the model is trained on K-1 folds of the data and evaluated on the remaining fold. This process is repeated K times, with each fold serving as the test set once. The results from each fold are then averaged to provide an estimate of the model's performance. After the Grid Search is complete, the best hyperparameters found are printed using the best_params_ attribute of the grid_search object. The best estimator found by Grid Search is also used to predict the target values for the test set, and the model's performance is evaluated using mean squared error (MSE) and R-squared score. This process is repeated for each hyperparameter to find the best combination of hyperparameters that gives the highest R-squared score. Once the best combination of hyperparameters is found, the model is retrained using the entire dataset with the best hyperparameters found. Working with the Hyperparameters proved to be tricky, with the first iteration scoring lower than the random Forrest model without tuning, and the last hyperparameter model which took a significantly longer amount of time to compute scored slightly lower as well. We are then led to believe the parameters are overtrained, and thus over fit the data. Testing with Ridge regression also proved to be worse than random forest but better than linear regression.

Results and Evaluation:

```
Linear Regression - R-squared score: 0.72
Random Forest Regression - R-squared score: 0.89
R-squared score with Ridge regularization: 0.74 (+/- 0.06)
```

The results indicate that the random forest regression model outperformed the linear regression model in terms of mean squared error and R-squared score. The R-squared score of the linear regression model was 0.72, while that of the random forest regression model was 0.89. This suggests that the random forest regression model is a better fit for the data.

Three rounds of hyperparameter tuning were performed on the random forest regression model. The first test R-squared score was 0.84. However, when the model was evaluated using the best hyperparameters, the R-squared score decreased to 0.88. This indicates that the performance of the model worsened with hyperparameter tuning.

Further evaluation of the models was conducted using Ridge regularization. The R-squared score with Ridge regularization was 0.74 (+/- 0.06). These performance metrics suggest that the linear regression model is not a good fit for the data and could benefit from further optimization.

Concluding remarks

The data we chose is from a car database that includes information about cars sold in the United Kingdom. The data set includes variables such as make, model, year, horsepower, and fuel type. The data has been cleaned by handling missing values, removing outliers, and dealing with inconsistencies in the data as well as updated to be easier to work with. Feature engineering has been performed to create new variables, such as the inflation rate, to improve the accuracy of the machine learning model. Feature selection was carried out by using correlation analysis, pair plots, and random forest regressor, and the SelectKBest function was used to select the most relevant features. The data was split into a training and testing set, and the random forest regression model was selected as the best model to predict the price inflated variable.

The work done on this dataset has given us valuable information and hands-on experience with working with Machine Learning. The process of data collection and cleaning is crucial in preparing data for machine learning. The feature engineering and feature selection techniques used in this project show the importance of creating new variables and selecting only the most relevant features to improve the accuracy of machine learning models. The findings also show that increasing the size of the testing set can lead to a more reliable estimate of the model's performance.

Future work could involve exploring other machine learning models or experimenting with different feature engineering and selection techniques to further improve the accuracy of the

models. Additionally, the dataset could be expanded to include more variables and more countries to create a more diverse and comprehensive dataset.