Machine Learning Assignment

**Introduction**

This report is the documentation of every step taken, definitions and algorithms used to build the model that accurately predicts tax revenues for upcoming fiscal years.

The problem statement

Revenue Services Lesotho (RSL) is the government agency responsible for tax administration and revenue collection in the Kingdom of Lesotho on which its primary objective is to ensure compliance with tax laws and regulations while facilitating economic growth and development through effective revenue mobilization.

As such, RSL needs to be ahead on forecasting what could possible happen in the upcoming years. For it to be precise, there have to be the implementation of the model based on historical data.

Project scope

This project covers only the implementation of the model to make predictions having being given the historical data. The building of the model will include all the necessary steps like data preprocessing, feature engineering, the model training and tuning.

Model development

The report will details all the taken in building the model based on the task given.

Task 1: data gathering and cleaning

Data gathering is collecting of useful information for a particular domain. In this case, the already available dataset namely “Data RSL” was given.

Data cleaning is the process of finding possible happened errors and anomalies that are available in the dataset. This includes the following: identifying missing values and anomalies or outliers.

What are missing values?

Missing values are specific variables with null or unknown data points in the given dataset.

Handling missing values

The missing values can be deleted, that is, rows and columns consisting of missing values can be removed.

Another method is imputation, which is replacing missing values with estimates. Imputation has its method namely; mean/median/mode imputation, k-nearest neighbors and model based imputation.

In our case, the given dataset contained missing values in this columns: reference period code, Reference period, Flags and flags code. We used arbitrary value of “0.0” to fill missing values.

Dealing with outliers or anomalies

Outliers are data points which gives unusual difference among majority of observations.

Box and scatter plot as visual methods can be used to detect outliers. Z-scores, Interquartile range and Distance metrics are statistical methods used to identify outliers.

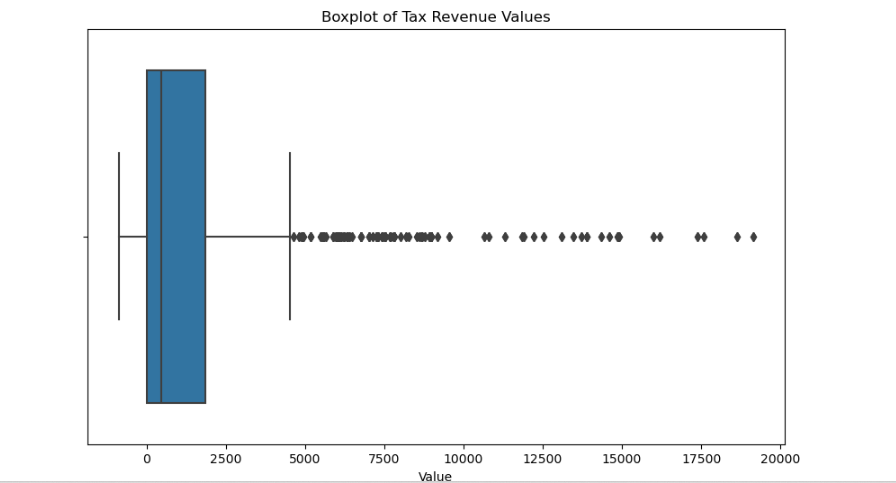
We used the Box plot to make visualization of outliers and Z-scores to make statistical analyzes of outliers.

Task 2: Exploratory Data Analysis

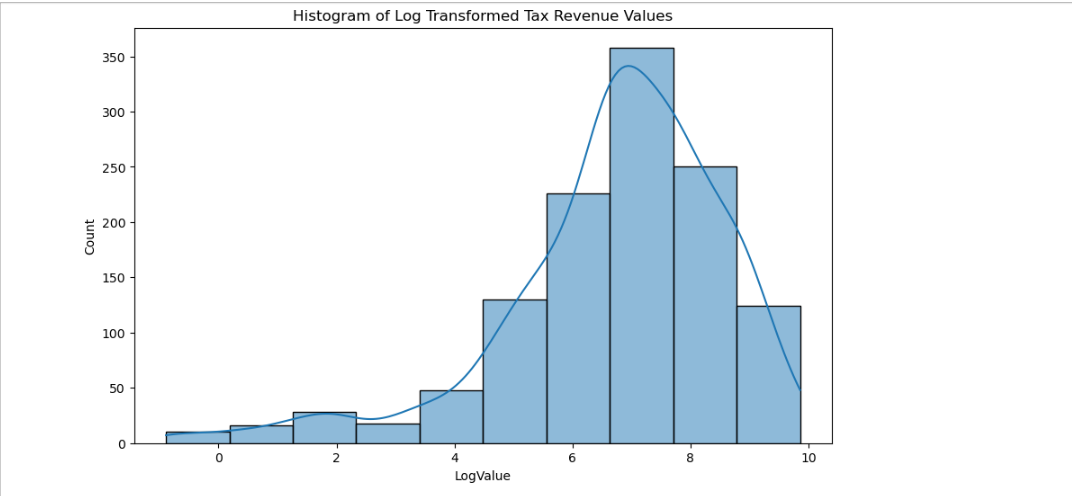
This is the graphical representation of information and data by the use of visual elements such as graphs, maps and charts. This help to understand relationships, patterns and anomalies like making use of scatter plots which is useful in revealing relationships between variables.

Data visualization is also important in feature selection whereby the use of matrices like correlation can help in selecting relevant features which are highly correlated with target variable.

Also in model evaluation, visual tools helps in evaluating model performance by comparing predicted values to actual values.

The figure below shows the boxplot of tax revenue values when detecting outliers.  


The picture below shows the histogram of log transformed tax revenue values after reducing the impact of outliers using Log transformation.

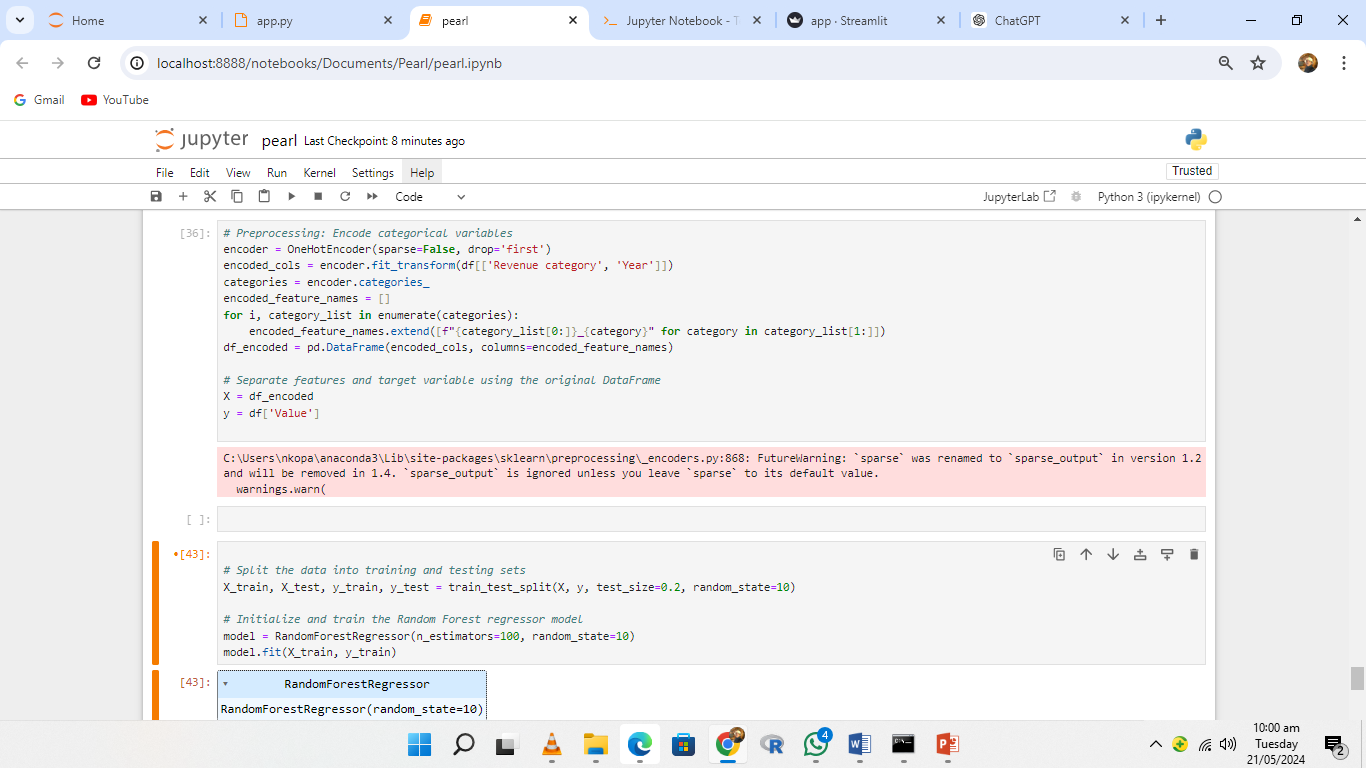


Task 3: Feature Engineering

This is the process of using domain knowledge to extract new features from the raw data in order to make machine learning algorithm works better. Steps in feature engineering includes: encoding categorical variables, scaling and normalization, creating new features and feature selection.

In our scenario, for machine learning algorithm to perform efficiently; we used one-hot encoding to convert categorical values in the column revenue category to binary. This allows the algorithm to understand the variables because the learning algorithm interpret the format of 0 and 1s only.

To create new features, we used scikit-learn one-hot encoder to transform new categorical values.



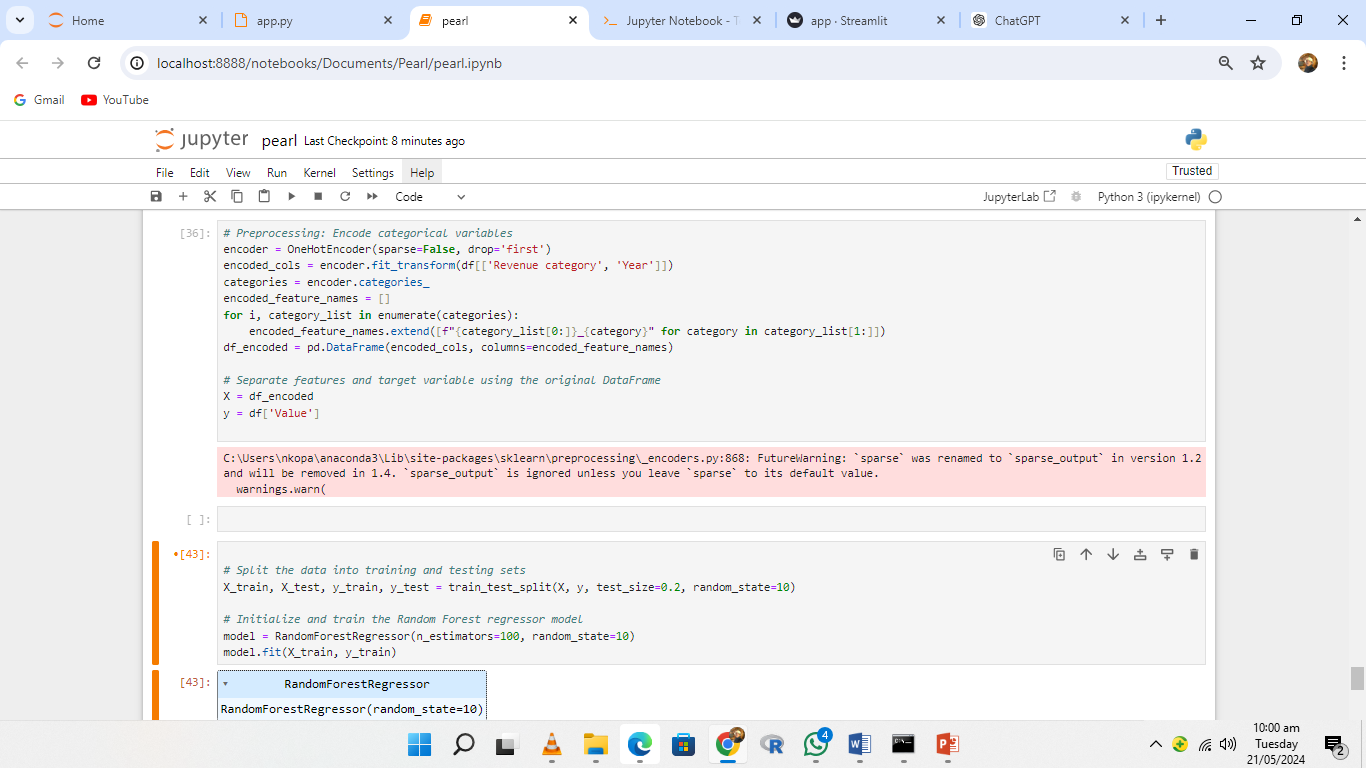
Task 4: Model training

The process by which a machine learning algorithm is fed data and adjusts its parameters to learn the underlying patterns or relationships within the data. This process involves optimizing the algorithm's parameters to minimize errors and improve its predictions on unseen data.

Before heading to model training, more importantly the data have to be split into training and testing sets on which there must be specified test size. We used the test size of float number 0.2 which indicates that 20% of the data will be used to make prediction of the unseen data whereas the remaining 80 % will be used to train the model. There is also random state which controls the shuffling and splitting of data. We used random state integer of 10, this means that running the same code multiple times will generate different split.

Model training with Random Forest Regressor

We decided to make utilization of this model because we are working with dataset which contains continuous values. Random Forest Regressor consist of decision trees that has internal node which presents decision based on the feature. In this case, the extracted features from revenue category will be presented through the model by decision trees.



**GROUP MEMBERS**

1. **NKOPANE PONDO 901011001**