Tax Prediction Model

Machine Learning – Artificial Intelligence

Sello Majoang Mokoena Nyamatana Lebusetsa Moorosi Malebabo Moorosi

Group project

**Problem statement**

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

Collecting revenue and tax administration is a crucial process in funding public services, infrastructure development, and socioeconomic programs in Lesotho. RSL’s efforts contribute to the country's fiscal sustainability and overall economic stability.

A task has been placed in our hands to develop a machine learning model that accurately predicts tax values from the given historical data.

**Data gathering and cleaning**

Data cleaning is crucial step in the machine learning pipeline, this process in concerned with identifying and removing missing values, duplicate or irrelevant data. The aim is to provide assurance the training data is accurate, consistent and free of errors, as incorrect or inconsistent data can negatively impact the performance of the machine learning model [1]. The dataset that has been provided had missing values, so the null values have been removed from the dataset to prevent the model from being biased and that the predictions should not be skewed. Some of the columns were dropped as they were not applicable for our machine learning model leaving only revenue category, year and tax value. Revenue category and year being the independent values and tax value being the dependent value as it our target of prediction. The predictors had duplicated values which they too were also removed from the dataframe which was created when importing the comma separated values file.

**Exploratory Data Analysis**

Exploratory data analysis involves analysing and creating visual displays to better understand its key characteristics, uncover patterns, identify relationships between variables. The dataset was explored to learn about it, it was explored by visualizing the data so that we can have a visual of the data to better understand it. Various visualization charts were employed for this purpose.

* Visualization of the distribution of the value was made to learn the frequency of each value in the dataset. Using seaborn library to plot the histogram: this was accomplished. The KDE curve provides a smoothed estimate of the distribution, helping to visualize the underlying pattern more clearly.
* ##outliers
* We did statistical analysis, statistical analysis is a process of understanding how variables in a dataset relate to each other and how those relationships depend on other variables. Visualization can be a core component of this process because, when data are visualized properly, the human visual system can see trends and patterns that indicate a relationship [2]. We used a scatter plot to visualize the relationship between the year and tax value.
* Understanding how numerical values relate to each other in a dataset is really important for making good decisions [3]. Using a correlation matrix and heatmap, like we did in the code, helps us see these relationships visually. It's like putting numbers on a map to see where they're close or far from each other. This helps us understand how different numbers in our data are connected. The colours in the heatmap, like red and blue, show us if the connection between numbers is strong or weak, and if it's positive or negative. Also, the numbers we add on the heatmap give us exact details about how strong the connections are between pairs of numbers. This way of analysing data is well-trusted in science, which makes our findings more reliable and solid (outliers)

**Feature Engineering**

**Model Training**

**Model Tuning**

**Prediction Format**

After the development of the machine learning model, a user interface was developed to allow for user interaction with the model. This interface was developed using HTML, CSS, Javascript together with Nodejs for providing an application programming interface (API) for facilitating communication between the model and the user. The interface allows for both selection of revenue category and year to predict the tax value.

**Conclusion**

**References**

[1] Geeks for Geeks, "Machine Learning Overview," Geeks for Geeks. [Online]. Available: https://www.geeksforgeeks.org/machine-learning/. [Accessed: May 22, 2024].

[2] Wilkinson, L., Friendly, M., & Reynolds, M. (2009). The grammar of graphics (2nd ed.). Springer.

[3] Chong, G. Jian, M. Zhiying, J. Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. Volume 4, Issue 4, December 2020.