IT5006 Fundamentals of Data Analytics

Analytical Approach for Equitable Taxation Model

(Data Scientist Track)

Team Members (Team 17)

Arvind Subramanian A0228522J e0674506@u.nus.edu
Ramesht Shukla A0228506E e0674490@u.nus.edu
Supratik Sekhar Bhattacharya A0228511M e0674495@u.nus.edu

Introduction

Problem Statement: Inequitable taxation increases the burden on families that need help.

Primarily, there are three types of tax policies: progressive, regressive, and proportional, and most countries around the world adopt a progressive taxation system. A percentage of income forms the tax, the quantum of which may vary according to income thresholds and may also include a fixed component. The general principle is that those who earn more pay more. But this leads to economic inequality globally and leads to various concerns as it impacts health and social welfare. Finding an optimal tax policy that optimizes inequality and productivity is an unsolved problem [1].

Though this seems fair on the outset, such a system does not account for the differences in spending needs or habits of various households. Due to ailments, families with more dependents or higher medical costs would incur higher subsistence costs. However, they would be charged the same amount of tax as another household that earns the same amount but has fewer subsistence costs. It discourages individuals from working, which leads to lower productivity.

There are mechanisms for tax relief for cases where you have more dependents than other households, medical costs, donations to charity, and many more. However, the scope of these tax reliefs is generally limited and would not cover other circumstances which may be just as valid. Even when one is eligible for tax relief, obtaining the relief is generally not automated and tedious. It also opens the possibility of fraudulent claims or declarations.

We aim to address these issues by augmenting the existing taxation system with a predictive model. It would allow us to predict the expenditure of a particular household for the past year and automatically adjust the tax levied on them.

To build this model, we needed the data on the characteristics of a household and its income. It includes age, education level, occupation, and the number of members, to name a few. Thus, we decided on the Philippines Household Income and Expenditure dataset. This dataset contains information on 41,544 households in the Philippines across 60 attributes.

Fig 1: Income vs. Family Size in the Philippines.

A quick exploration of our dataset confirms our initial analysis. As shown in Fig 1 above, income does rise marginally as the number of family members increases. However, this increase is not proportional. Particularly in the family size of 4-8, the average household income is almost the same. In some cases, families with more members even make less income. It shows that though they may have more mouths to feed, some families do not make sufficient extra money to offset the cost. However, they are taxed the same amount as those with smaller families and seek to make taxation more equitable.

Also, the Philippines government census board collated and presented this data. To curate this dataset, they employed the "Complete Enumeration" approach [2]. Utilizing such information, which government census agencies commonly collect, would make our model more applicable worldwide.

Data Cleaning

1. Data Preprocessing

Removal of Lower Income Households. As per the taxation policy enforced by the Philippines Bureau of Internal Revenue shown in Fig 2 below, households earning below 250,000 Philippine Pesos (PHP) annually do not need to pay income tax. As such, we can remove these households from the data frame.

Amount of Net Taxable Income		Rate
Over	But Not Over	
-	P250,000	0%
P250,000	P400,000	20% of the excess over P250,000
P400,000	P800,000	P30,000 + 25% of the excess over P400,000
P800,000	P2,000,000	P130,000 + 30% of the excess over P800,000
P2,000,000	P8,000,000	P490,000 + 32% of the excess over P2,000,000
P8,000,000		P2,410,000 + 35% of the excess over P8,000,000

Fig 2: Different taxation tiers followed by the Philippines government.

After the removal, the dataset shape changes to 12807 x 60. We can conclude that many households in this dataset are below the taxation threshold.

Managing Missing Values. Only 2 features have missing data, each with 3114 null values (Household Head Occupation and Household Head Class of Worker). These values were object values. We did not want to remove them as they may hold statistical significance. We are trying to develop a new taxation system for our problem statement that the Household Head's occupation would significantly influence. Hence, we tried to estimate and impute these missing values.

Assuming that the combination of age, gender, and educational qualifications would affect one's occupation, we can examine the correlation of 'Household Head Occupation/Class of Worker' with other features having the prefix "Household Head...". It can be used to estimate the missing values. However, upon closer examination of the "Household Head Job or Business Indicator" feature, we notice that the 3114 "null" values occur because they were unemployed. Hence, we replaced these "null" values with "Unemployed" as a categorical variable while still holding the analytical value.

```
data.isnull().sum()
```

```
TotalHouseholdIncome                0
Region                              0
TotalFoodExpenditure                0
MainSourceofIncome                  0
HouseholdHeadOccupation          3114
HouseholdHeadClassofWorker       3114
TypeofHousehold                     0
```

Fig 3: Missing values in data.

```
data['HouseholdHeadJoborBusinessIndicator'].value_counts()
```

```
With Job/Business    9693
No Job/Business      3114
Name: HouseholdHeadJoborBusinessIndicator, dtype: int64
```

Fig 4: Data values to be imputed.

Variable Identification. Several features describe the spending habits of each household. For example, total food expenditure, bread, and cereal expenditure, education expenditure, and many more.

```
expenselist = []
for i in data.columns:
    if 'Expenditure' in i:
        expenselist.append(i)
expenselist
```

```
['TotalFoodExpenditure',
 'BreadandCerealsExpenditure',
 'TotalRiceExpenditure',
 'MeatExpenditure',
 'TotalFishandmarineproductsExpenditure',
 'FruitExpenditure',
 'VegetablesExpenditure',
 'RestaurantandhotelsExpenditure',
 'AlcoholicBeveragesExpenditure',
 'TobaccoExpenditure',
 'ClothingFootwearandOtherWearExpenditure',
 'HousingandwaterExpenditure',
 'MedicalCareExpenditure',
 'TransportationExpenditure',
 'CommunicationExpenditure',
 'EducationExpenditure',
 'MiscellaneousGoodsandServicesExpenditure',
 'SpecialOccasionsExpenditure',
 'TotalNecessaryExpenditure',
 'TotalUnnecessaryExpenditure']
```

Fig 5: Expense-related features in the dataset.

We can further segregate this to understand how much a household spends on essentials vs. luxuries. It could offer a valuable discriminating factor for us to adjust taxation. We predict the expenditure values (dependent variables) based on the other non-expense related features (independent variables). The most transparent methodology to attempt would be a multivariate analysis due to the number of features available.

Categorical Variable Conversion. There are many categorical values in the dataset that needed to be dummy encoded to be independent variables.

Feature Selection. There are a total of 60 features, many of which seem to be redundant. All the features related to expenditure can be merged into essentials and non-essentials. For example, bread and cereal expenditure, education expenditure can be counted as a necessary expenditure. Features like alcohol expenditure and tobacco expenditure come under unnecessary expenditure.

Note that some features have been excluded from this as they are redundant or ambiguous. Ambiguous in this context means that they are likely to contain both scenarios of necessity and luxury.

- 'Total Food Expenditure' is equivalent to adding up the spending on individual food items.
- 'Housing and Water Expenditure' is ambiguous as it could mean rent for a small apartment or mortgage for a luxury apartment.
- 'Education Expenditure' should be encouraged, but there could also be people who splurge on expensive private schools or overseas education.
- 'Crop Farming and Gardening Expenses' may not apply to enough people above the minimum taxable income bracket to provide meaningful analysis.

We also do not need so much granularity on the number of appliances and assets owned. Hence, we combine them under a single feature named 'Total Movable Assets.'

```
data['TotalNecessaryExpenditure'] = data['BreadandCerealsExpenditure'] + data['TotalRiceExpenditure']
data['TotalUnnecessaryExpenditure'] = data['RestaurantandhotelsExpenditure'] + data['AlcoholicBeverage
data['TotalMovableAssets'] = data['NumberofTelevision'] + data['NumberofCDVCDDVD'] + data['NumberofCom
```

Fig 6: Consolidation of features.

Feature Engineering. We can generate new features by extracting useful information from the existing features. These new features would be more valuable for modeling than the individual features themselves. We created new features that relate to the expenses and quality of life of each household.

```
data['Dependants'] = data['TotalNumberofFamilymembers']-data['Totalnumberoffamilymembersemployed']
data['HouseSpacePerPerson'] = data['HouseFloorArea']/data['TotalNumberofFamilymembers']
data['MovableAssetsPerPerson'] = data['TotalMovableAssets']/data['TotalNumberofFamilymembers']
```

Fig 7: New features that provide more precise interpretation.

We created three new features, as seen in Fig 5. The "Dependents" feature was created by calculating the number of unemployed family members in the household. We deduced the "HouseSpacePerPerson" by calculating the ratio between the total area of the house and the number of family members. Finally, "MovableAssetsPerPerson" was calculated by dividing total movable assets by total family members.

We also created a new column for the current taxes paid by each household, based on personal income tax quantum in the Philippines (as per Fig 2).

Finally, we created a new consolidated data frame combining the model's independent and dependent features.

```
mydata.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12807 entries, 0 to 41532
Data columns (total 21 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   TotalHouseholdIncome                  12807 non-null  int64
 1   Region                                12807 non-null  int8
 2   MainSourceofIncome                    12807 non-null  int8
 3   AgriculturalHouseholdindicator        12807 non-null  int64
 4   ImputedHouseRentalValue               12807 non-null  int64
 5   HouseholdHeadSex                      12807 non-null  int8
 6   HouseholdHeadAge                      12807 non-null  int64
 7   HouseholdHeadMaritalStatus            12807 non-null  int8
 8   HouseholdHeadHighestGradeCompleted    12807 non-null  int8
 9   HouseholdHeadJoborBusinessIndicator   12807 non-null  int8
 10  HouseholdHeadOccupation               12807 non-null  int16
 11  HouseholdHeadClassofWorker            12807 non-null  int8
 12  TypeofHousehold                       12807 non-null  int8
 13  Electricity                           12807 non-null  int64
 14  MainSourceofWaterSupply               12807 non-null  int8
 15  Dependants                            12807 non-null  int64
 16  HouseSpacePerPerson                   12807 non-null  float64
 17  MovableAssetsPerPerson                12807 non-null  float64
 18  TotalNecessaryExpenditure             12807 non-null  int64
 19  TotalUnnecessaryExpenditure           12807 non-null  int64
 20  CurrentTaxPaid                        12807 non-null  float64
dtypes: float64(3), int16(1), int64(8), int8(9)
memory usage: 1.9 MB
```

Fig 8: New dataset created by combining independent and dependent variables.

Outlier Removal. As shown in Fig 7a, there are significant outliers in the expenditure data. It is from a small handful of households that spend abnormally high amounts. It would be misrepresentative of the majority, which may be relatively frugal. If we do not remove these outliers, our model would be biased towards predicting higher values for expenditure.

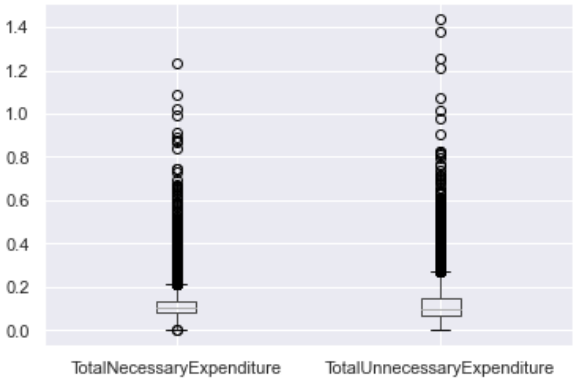

Fig 9a: Boxplot showing the outliers in the dataset.

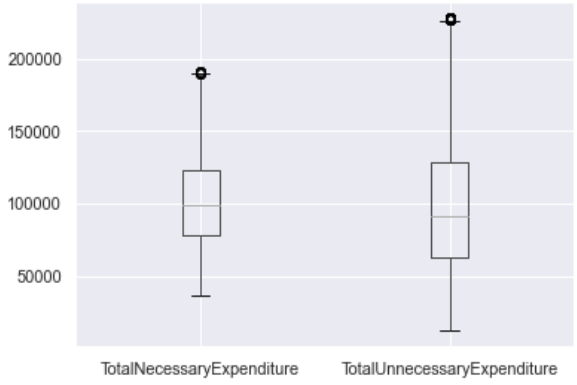

Fig 9b: Boxplot showing the dataset without outliers.

As observed, in Fig 7b with a Z-score threshold of 1.25 eliminates most of the outliers. The number of data points has decreased only slightly to 10869, and the size of the remaining dataset is still significant. Finally, we believe that this dataset should provide a more accurate prediction model.

Methodology

We choose regression to develop our model as we are looking to predict a value based on multiple known features.

Splitting of Dataset. We split the dataset into training and test sets with the ratio of 80:20, where training data formed 80% while the remaining 20% formed the testing dataset.

Modeling. We employed different models for our prediction, namely: Ridge Regression, Multiple Linear Regression, Ridge Regression with Polynomial Regression, and Multiple Linear Regression with OLS. Table 1 shows the R- Score for each of the models.

Table 1: R-Scores of Various Models

Models	R-Score
Multiple Linear Regression	0.234
Ridge Regression	0.234
Ridge Regression with Polynomial Regression	0.286
Multiple Linear Regression with OLS	0.929

Multiple Linear Regression with OLS, by far, outperforms other regression models. It was, therefore, the model of choice.

Prediction. We predicted two separate dependent variables: 'Predicted Necessary Expenditure' and 'Predicted Unnecessary Expenditure.'

Tax Adjustment. The tax adjustment was calculated based on the predicted expenditure values. The equation for adjusted tax is as follows:

$$AdjustedTax = \frac{UnnecessaryExp}{NecessaryExp}\ CurrentTaxPaid$$

With this algorithm, the higher the unnecessary spending w.r.t necessary, the higher the tax.

The actual and predicted values and current and adjusted taxes were consolidated into a new data frame.

Results. The results indicate that the recommended taxation levels are more equitable than existing taxation schemes.

As shown in Fig 10a, there is one fixed tax payable in the existing taxation system for a particular income level, regardless of expenses. However, with the adjusted tax we have recommended, taxation varies across an entire spectrum (denoted by the red box) according to the expenditure.

As shown in Fig 11 below, the adjusted tax is higher than the current tax with increasing unnecessary expenditure rates. It makes the taxation more equitable whereby unnecessary expenses are deemed luxuries and, therefore, taxed more to provide tax relief for those with greater subsistence expenses.

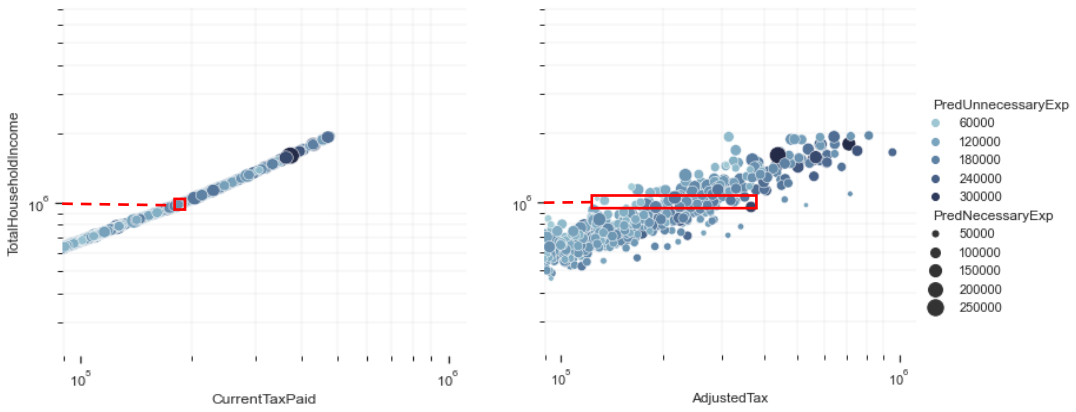

Fig 10: (a) Fixed tax amount for a certain income. (b) Tax amount varies across an extensive range as per the red box.

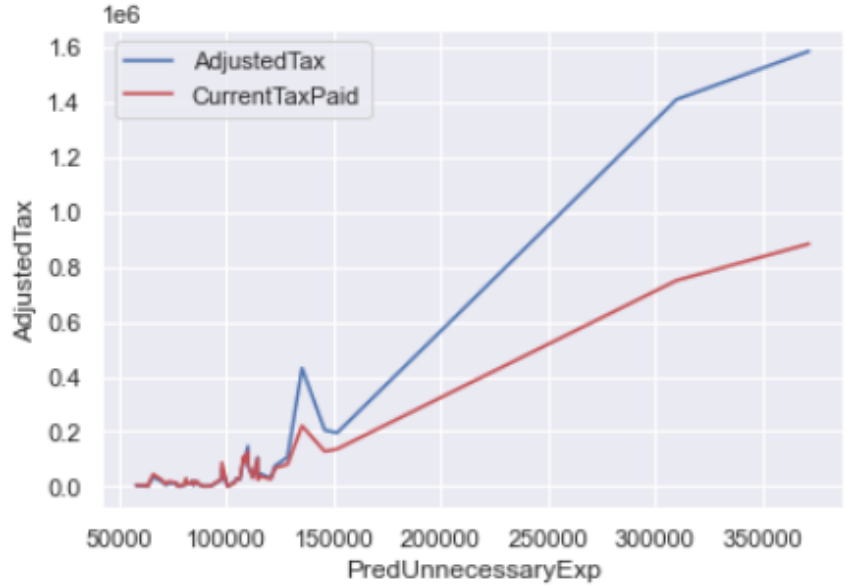

Fig 11: Adjusted Tax higher than Current Tax for unnecessary expenditure.

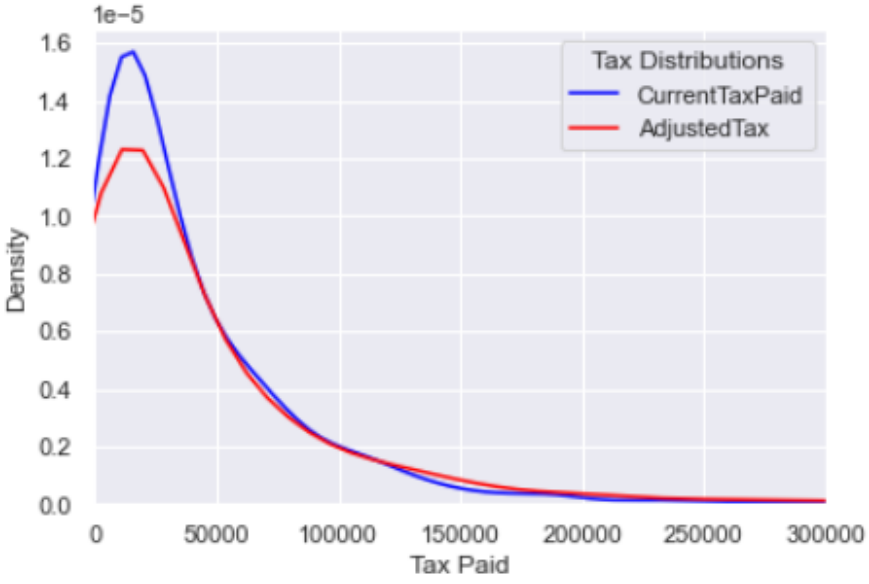

Fig 12: Distribution of tax amounts.

As shown in Fig 12, at the lower-middle tax brackets (PHP 0-300000), fewer households are paying lower taxes under the adjusted tax system. Finally, we can see an increase of 13.4% in tax revenue collected with the adjusted tax as opposed to the current tax system.

Our model intends to offset tax relief for subsistence spending using a higher tax on luxuries. Therefore, the significant increase in tax revenue suggests that a higher number of households tend to spend wastefully than needy households! Simultaneously, it proves that while being more equitable with taxation, our adjusted tax model does not shortchange the government on its tax revenue.

Conclusion

The Equitable Tax Model that we have created effectively addresses the issues mentioned in the problem statement. Based on the pre-collected information about each household, governments can accurately predict their necessary and unnecessary expenditure. The model has an R-score of 0.929 in predicting necessary expenses and 0.877 in predicting unnecessary expenses.

We can achieve more equitable taxation by using a formula that uses a ratio of unnecessary/necessary expenses to adjust the current tax. Naturally, having higher necessary expenses would therefore lead to lower taxes and vice versa.

The model we have created does not replace the existing tax system but instead augments it. The only modification to the actual tax amount is through the abovementioned ratio. Therefore, it would be easier for governments to understand and implement.

Our model also uses common census data to make its predictions, like the number of dependents/education levels. All countries usually collect this data about their citizens. As such, any country could apply our model.

There are a few enhancements that can still be made to the model. Before implementation, more expert opinions could be sought to advise on the features to select/additional features to gather, likely to have the highest impact on expenses. The tax adjustment ratio could also be customized to increase/decrease weightage on additional spending. It would help to further refine the proposed tax.

In conclusion, our model is based on sound socio-economic principles and addresses the problem statement effectively.

References

1. Zheng, Stephan, et al. “The ai economist: Improving equality and productivity with ai-driven tax policies.” arXiv preprint arXiv:2004.13332 (2020).
2. https://psa.gov.ph/article/technical-notes-family-income-and-expenditure-survey-fies
3. https://www.bir.gov.ph/index.php/tax-information/income-tax.html