MS984: Data Analytics in Practice

Case Study 3: Gross Annual Income Estimation and Recognising Limitations

Group 4

Le Phuong Linh | Shrishti Sridhar Manja | Hrutuja Mangesh Patkar |

Blaise Marvin Rusoke | Liyi Tan

1. Introduction

Gross annual income values are vital to credit assessment service providers in the UK, including Experian. This data is used as the metric through which the creditworthiness of individuals is evaluated prior to mortgage application approvals/denials. However, accessing the gross annual income data is not feasible in reality. Thus, the report introduces a mathematical algorithm developed using systematic programming techniques that accurately estimates the gross annual income from net monthly income. The developed solution caters for individuals in both Scotland and other regions in the UK and uses the appropriate national insurance and income tax brackets. The report presents the methodology in detail, discusses the solution’s strengths and limitations, and advocates for a comprehensive approach to estimating the gross annual income.

2. Data Exploration

The dataset utilised in this analysis is a representative sample of the United Kingdom's population. It comprises two subsets: a test dataset and a holdout dataset. The training dataset contains three key columns: 'Location,' representing the region of the UK; 'monthly_net_income,' denoting the monthly net income of individuals; and 'annual_gross_income,' indicating the annual gross income of the individuals. The 'Location' column has two distinct values: 'sco,' an abbreviation for Scotland and 'rou,' an acronym for the rest of the UK. The test dataset shares the same structure as the training dataset but the 'annual_gross_income' column needs to be filled. The primary objective of this study is to develop a precise mathematical model predicting the annual gross income of individuals based on the available 'monthly_net_income' data.

Given the context from which the data is derived, all values should be positive. Hence, an initial exploration was conducted to identify and address any missing or negative values. In the training dataset, there were no NaN values, but five entries with negative values were identified. To maintain data integrity, these instances were filtered out. A new feature representing annual net income was also introduced by multiplying the 'monthly_net_income' values by 12, offering a convenient representation of the annual net income to be used in the model.

3. Methodology and Results

Developing the function to convert from net to gross income includes establishing net income brackets and formulating the equation to compute gross income for each range. The rationale is to determine the specific bracket to which the provided net income belongs and subsequently apply the corresponding formula to determine the gross income.

The upper limit for each gross income range is used to determine the upper limit for the net income range, by subtracting the maximum tax, insurance, and personal allowance. For instance, with the range from 11904 to 12570, the net lower limit is unchanged (11904) as no tax and insurance is applied. The net upper limit is computed as follows:

*12570 - National Insurance Tax - Income Tax = 12570 - (12570-11904) * 0.1325 - 0 = 12481.76*.

Thus, the net income range is from 11904 to 12481.76. The same principle is applied to determine the remaining net income ranges.

The formula is adjusted from the gross to net function to generate the gross income. With the above example, the annual net income is calculated through the formula:

*annual_net_income = gross_income - ((gross_income-11904) *0.1325).*

This formula can be reversed to predict the gross income as the subject:

*gross_income = (annual_net_income - 11904*0.1325)/ (1-0.1325).*

To visually analyse potential patterns of errors across gross income values, a scatterplot of annual gross income against the difference between predicted and actual annual gross income values was plotted.

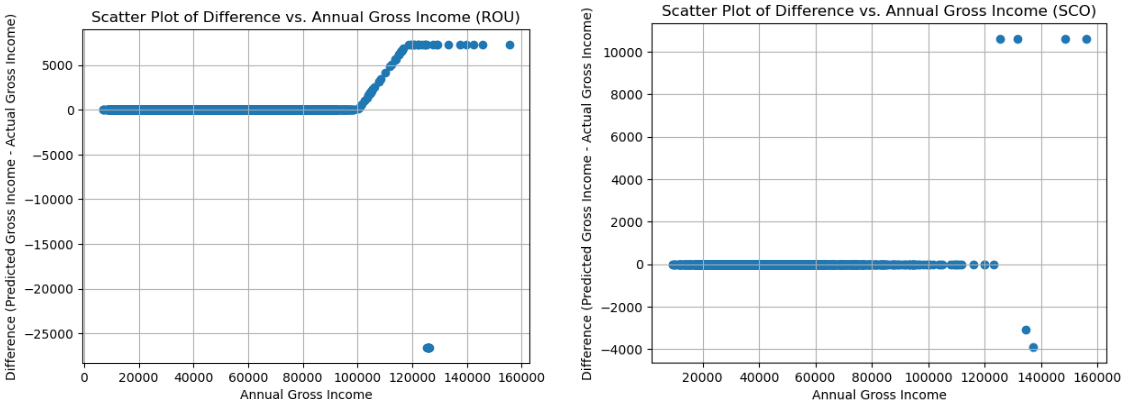

Figure 1: Scatterplots of model error against annual gross income

These functions yielded accurate outcomes for gross incomes below £100,000 and £125,140 on ROU and SCO test datasets respectively. Beyond £100,000, the personal allowance is reduced, and the reduced amount was subjected to a tax rate of 42% or higher for Scotland and 40% or higher for the UK. However, for gross incomes surpassing the thresholds, the differences between the predicted and provided annual gross income are quite high. This situation may be attributed to the inconsistent rate applied on allowance

The model's margin of error on the ROU and SCO test sets are 0.26% and 0.11% respectively. Overall, the models margin error is 0.23%. A lower margin of error implies that the average prediction error, represented by MAE, is relatively closer compared to the average actual income. This makes sense since the model does not perform well on a larger range of values for UK tax systems as compared to Scotland's. Another reason could also be due to an overall lower number of annual gross income values over the range of 125,140 in Scotland's test set compared to ROU's, skewing the margin of error in favour of Scotland.

4. Strengths and Limitations

The developed algorithm makes use of systematic programming and takes advantage of well-thought-out Python functions to reverse the original gross-to-net calculation to the required net-to-gross computation. Because of its systematic and transparent nature, the algorithm is relatively easy to understand, debug, and customise as it is written in a clear 'Pythonic' form with a top-down approach.

However, for a real-world use case, a more robust technique would be required for this task. Real-world regression datasets are usually more problematic and cannot be solved by systematic programming functions but rather by more complex regression analysis methods which minimise the error between the predictions and true values.

Furthermore, because tax brackets may change throughout the year, the developed systematic programming function would need to be manually modified and its accuracy evaluated on the new tax brackets. This is rather a tedious process. In the event that the client decides to consider new features to compute the gross income such as loan repayments, the logic would also have to be manually added to the systematic function, which would keep expanding, and eventually become unmaintainable as more features are added.

Using systematic programming functions is not a sustainable approach. Regression analysis algorithms are better suited for such tasks and overcome all the limitations of the systematic programming approach highlighted in this section.

The dataset utilised for this analysis also only considers income tax and national insurance tax as the only deductions from an individual's gross income. A more reliable and robust approach for estimating gross annual income from net monthly income would include more transactions such as pension payments and student loans. There would also be a need to cater for unique sets of individuals such as self-employed individuals when conducting a substantial analysis. The fact that these essential items are not considered in our analysis greatly hinders the practicality of our solution.

Given more time, we would have developed and presented a more robust and reusable solution, that makes use of linear regression techniques to understand the underlying patterns in the data and make more accurate predictions. We also would have sought to use a richer dataset with more features to account for real-world gross income deductions and tailored our solution to also account for unique sets of individuals such as Self-employed persons.

5. Conclusion

In conclusion, the report proposes a mathematical function for estimating gross annual income from net monthly income, incorporating national insurance and income tax deductions. While our systematic programming approach demonstrated transparency, and ease of customization, and yielded positive outcomes, we acknowledge the limitations of manual adjustments for changing tax brackets and potential challenges with new features. We advocate for a more sustainable and robust approach, specifically the use of regression analysis machine learning algorithms and richer transactional data with more real-world features such as loan repayments.