University of Strathclyde Business School

Department of Management Science

MS984: Data Analytics in Practice

Case 3 – Experian: Net to Gross Income

Group 3:

Olumuyiwa Olajuwon
Susan Dangana
Michael Patrick
Ademola Oladimeji
Muneeba Nawaz

METHODOLOGY

To analyse the given datasets, a well-considered technique was developed. We performed initial brainstorming meetings to define the attributes in the dataset and decide on the suitable methodologies for data analysis. The datasets, provided by Experian, were partitioned into a test dataset and a holdout dataset. The former included attributes such as location, monthly net income, and provided annual gross revenue, whilst the latter consisted of details regarding location, monthly net income, and an expected annual gross income.

DATA ANALYSIS:

The analysis began by concentrating on the test dataset. The aim was to utilise Python code to calculate the yearly net income based on the given monthly net income. Throughout this procedure, certain problems were detected, particularly the presence of negative values in certain monthly net revenues. Exploratory data analysis was conducted to discover these inaccuracies.

The research was subsequently expanded to include the holdout dataset, in which Python code was employed to calculate the projected annual gross income. A customised function was created to calculate the yearly total revenue for the holdout CSV file.

RESULTS:

The comparative examination of the generated results and the delivered monthly net income for the test dataset, revealed a variance of less than 10%.

Moving to the holdout dataset, the Python code populated the annual gross revenue. Nevertheless, it was observed that there was an approximate margin of error of 10% when comparing the two codes for net to gross income and vice versa.

LIMITATIONS:

The analysis experienced a notable constraint because of the inadequate amount of data, which was limited solely to the attributes of location and monthly net income. The limited dataset presented difficulties due to its lack of detailed information, which is crucial for a thorough study. To overcome this limitation, the focus was directed towards optimising the existing data by employing techniques such as exploratory data analysis. The objective was to extract valuable insights within the given limitations. Nevertheless, the lack of essential factors underscored the necessity for careful interpretation, clear explanation of limitations, and an awareness of potential biases. This underscores the significance of recognising the limited scope when making conclusions based on the research.

IF GIVEN MORE TIME/DATA:

With more time and data, the existing methodology might be developed to improve analysis depth and robustness. Primarily, add attributes to the dataset to better understand the individuals or organisations under investigation. Demographics, employment history, and financial indicators including expenses, savings, and debt may be collected.

A more complex exploratory data analysis (EDA) could reveal patterns, correlations, and outliers in the enlarged dataset. Feature engineering could develop new variables that reveal participants' financial characteristics. Statistical modelling and machine learning algorithms can generate predictive models or reveal hidden trends, improving data interpretation.

A comparative analysis across dataset subsets, such as geographical regions or income levels, may reveal variances and patterns that a more limited analysis may miss. Financial behaviours may be revealed by longitudinal study of monthly net income.

Domain experts or stakeholders could be consulted to incorporate domain-specific expertise and improve data interpretation. Finally, documenting and sharing the enlarged approach and any new discoveries would improve analysis transparency and reproducibility.

CONCLUSION:

The developed methodology has proven to be helpful in managing the insufficient datasets. Using Python code, we calculated the annual gross revenue for the holdout dataset, even though there was a small margin of error. The results provide significant insights into the financial characteristics of the datasets.

RECOMMENDATIONS:

To improve the reliability of the analysis, it is advisable to continuously monitor and refine the Python code to minimise the occurrence of errors. Furthermore, it is advisable to incorporate further data validation measures to enhance the precision of the outcomes. To promote transparency and enable future analyses to be replicated, it is recommended to thoroughly record the process.