A Solution for Banking Institutions to Predict Clients' Income Level

Zuoru Jin – The University of Auckland
zjin526@aucklanduni.ac.nz

Abstract

This research paper delves into the intricate task faced by the Australia and New Zealand Banking Group Limited (ANZ) in predicting their customers' income levels. By leveraging data mining techniques and algorithms, it aims to enhance ANZ's marketing strategy and improve customer relationships. The methodology encompasses a thorough three iterations of data exploration, data preparation, data transformation, and the application of suitable data mining methods using three implementations. The study's primary findings suggest potential solutions for personalized banking services and more accurate credit assessments.

i. Introduction

The financial landscape is continually evolving, with banks seeking innovative methods to understand their customers better and tailor their services. This research paper endeavors to harness the power of data analytics to predict customers' income levels for ANZ, one of the largest banking institutions in the Oceania region. Using tools such as IBM SPSS Modeler, Python scikit-learn, and PySpark, this paper will analyze customer data and reveal unseen patterns.

The primary objectives of this research paper were:

1. To refine ANZ's marketing strategies using data-driven insights.
2. To strengthen customer relations by addressing their financial aspirations and needs comprehensively.
3. To craft personalized banking solutions aligned with individual customer income predictions.

By understanding the relationship between variables like marital status, education level, and capital gains with income, this research paper aims to provide ANZ with actionable insights that can potentially enhance customer experiences.

ii. Practical Problem

Modern banking institutions are increasingly recognizing the importance of advancing their marketing strategies and fostering deeper customer relationships through accurate prediction of their customers' income levels (Osipenko, D., 2016)[3]. By accomplishing this, banks can offer more personalized banking services and fine-tuned credit assessments that satisfy the unique financial needs and aspirations of each customer. Such a proactive approach not only enhances the value proposition for customers but also strengthens the trust and bond between the bank and its customers (Yoon, C., 2010)[4]. In an era where the financial landscape is rapidly evolving and competition is intense, personalization becomes a key differentiator. Banks that harness the power of data analytics to gain insights into their customers' financial behaviors and preferences can better position themselves to deliver customized solutions (Trihas, N. et al., 2013)[1]. Moreover, this shift towards a more individualized banking experience can lead to better financial products, more accurate risk assessments, and even innovative services that may not have been previously considered (Ching, R. et al., 2007)[2]. Ultimately, by understanding and predicting the income levels of their customers, ANZ can create a more flexible and adaptive business model that not only drives growth but also fosters long-term customer loyalty.

iii. Research Problem

In the era of personalized banking, understanding the financial background of a customer is of paramount importance (Zhang, X. et al., 2020)[5]. As banks increasingly shift towards offering personalized services, the ability to predict a customer's financial standing becomes invaluable (Hassaan, M. et al., 2023)[8]. This leads to the research question: Can we

accurately predict a customer's income bracket using their existing information?

Delving into this problem, this research seeks to assess the feasibility and precision of such predictions (Mogaji, E. et al., 2022)[7]. Unraveling this could significantly shape the way banks strategize their services, ensuring they are more aligned with individual customer needs (Altwaijri, A. S., n.d.)[6].

iv. Research Objectives

The goal of this research project is to fulfill the following objectives:

1. Marketing Strategy Refinement: Investigate methods to improve marketing strategies, emphasizing the inclusion of data analytics and customer income predictions for ANZ.
2. Strengthening Customer Relations: Explore solutions to enhance customer relationship management by understanding and addressing individual financial needs and aspirations.
3. Personalized Banking Solutions: Develop mechanisms to provide personalized banking solutions, focusing primarily on income predictions to ensure services are aligned with individual customer needs.

v. Literature Review

Introduction

The modern banking landscape has gone through a significant transformation, with a pronounced shift towards personalization and data-driven decision-making. This review delves into the existing literature to understand the importance of predicting customers' income levels, the role of data analytics in banking, and the implications for marketing strategies and customer relationship management.

The Importance of Personalized Banking and Customer Relationship Management

Trihas et al. (2013)[1] underscored the significance of efficient customer relationship management (CRM) in the tourism and hospitality sectors. While their study primarily focused on these sectors, the principles they highlighted are equally applicable to the banking industry. They emphasized the role of novel e-marketing strategies in fostering deeper customer relationships. In a similar situation, Yoon (2010)[4] highlighted the antecedents of customer satisfaction in online banking in China. The study found that a proactive approach, which caters to the unique financial needs and aspirations of each customer, not only enhances the value proposition but also strengthens the trust between the bank and its customers. This aligns with the first research objective, which emphasizes refining marketing strategies and the importance of CRM in banking.

Data Analytics and Predictive Modeling in Banking

The power of data analytics in predicting customer behavior and preferences is undeniable. Ching et al. (2007)[2] applied data classification techniques for churn prediction in retailing. Their findings suggest that similar techniques could be employed in the banking sector to predict various customer behaviors, including their income levels. Osipenko (2016)[3] directly addressed this by focusing on credit card interest and transactional income prediction using an account-level transition panel data model. The study highlighted the potential of advanced statistical techniques in accurately predicting customers' financial behaviors. This corresponds with the research problem, which seeks to assess the feasibility of predicting a customer's income bracket using their existing information.

Zhang et al. (2020)[5] further expanded on this by discussing personalized digital customer services for consumer banking call centers using neural networks. Their research emphasized the era of personalized banking and how crucial understanding the financial background of a customer is. This aligns with the second research objective, which aims to strengthen customer relations by addressing individual financial needs.

Cultural and Societal Implications in Banking

Altwaijri (2015)[6] conducted an empirical investigation into the acceptance of online banking in the context of Saudi national culture. The study emphasized the role of national culture in shaping marketing strategies and the acceptance of banking channels. This cultural perspective is crucial, especially when considering personalized banking solutions on diverse customer bases.

Hassaan et al. (2023)[8] explored the adoption of smart banking services in Pakistan, emphasizing the extended application of the UTAUT2 model. Their findings suggest that societal factors, such as the "big brother effect" and information privacy concerns, play a significant role in the adoption of banking services. This is particularly relevant to the third research objective, which focuses on the development of personalized banking solutions.

The Role of Artificial Intelligence in Financial Services Marketing

Mogaji et al. (2022)[7] provided a guest editorial on artificial intelligence (AI) in financial services marketing. They highlighted the transformative potential of AI in shaping marketing strategies and understanding customer behaviors. The insights from this editorial can be utilized in achieving the research objectives, especially in refining marketing strategies and developing mechanisms for personalized banking solutions.

Conclusion

The literature underscores the importance of personalized banking and the role of data analytics in shaping the future of the banking industry. By adopting the power of predictive modeling, advanced statistical techniques, and AI, banks can refine their marketing strategies,

strengthen customer relationships, and offer personalized solutions that cater to individual financial needs and aspirations.

vi. Research Methodology

This research adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) process methodology. Beginning with a comprehensive business understanding aims to identify and assess the objectives and current situation in the banking domain. The dataset, containing 32,562 records and 14 attributes, is then explored to understand its quality and characteristics. Through data preparation and transformation stages, it is cleaned, integrated, and optimized for subsequent mining. The appropriate data mining methods and algorithms are chosen to align with the objectives. In the data mining phase, multiple algorithms and models were used to extract related patterns and insights. Finally, the mined patterns are interpreted, visualized, and assessed for relevance and reliability. This approach ensures that the insights derived are both comprehensive and actionable when facing the challenges in the modern banking industry.

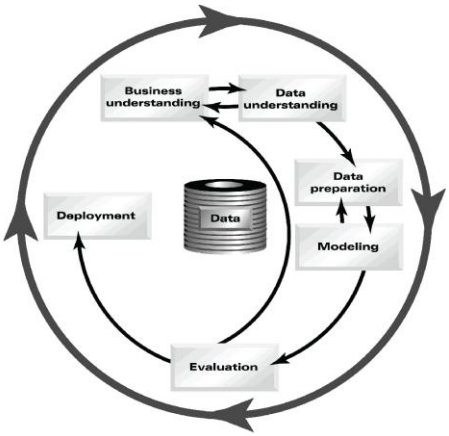


Figure 1: CRISP-DM Process (SPSS, 2007)

vii. Design of Processes

The process designed for this research is an integration of data science methodologies with modern tools and frameworks.

Tool Selection

The research utilized IBM SPSS Modeler, Python scikit-learn library and PySpark library for data preparation, data mining, and model building.

Data Pipeline

- Preprocessing: The preprocessing phase is the basis of this project, ensuring the quality of the dataset for mining.
    - Plausibility Check: An initial examination is conducted to ensure the integrity

of the data. While outliers and extreme values are often removed, care is taken not to exclude genuine or significant deviations.
        - Handling Missing Values: Rows with missing data points are excluded to maintain dataset consistency.
        - Outlier Management: Extreme values and outliers are prudently removed until the dataset reaches an acceptable state of balance and representation.
        - Feature Selection: Features are removed upon their relevance and ethical consideration.
        - Derived Attributes: New attributes are constructed based on the existing attributes. It provides more insights into the data-mining.
        - Dataset Balancing: To avoid model bias, efforts are made to ensure the dataset is balanced across different categories.
        - Dataset Integration: Different data sources or fragments, if any, are integrated into a unified dataset.
        - Reclassification: Attributes, especially categorical ones, are transformed using methods like one-hot encoding. Binary classifications are introduced to handle skewed attributes.
        - Redundancy Removal: To ensure the efficiency of subsequent processes, redundant attributes are identified and eliminated.
    - Data Mining: With a cleansed and ready dataset, the mining phase kicks off.
        - Dataset Partition: The dataset is divided into training and test sets using an 80%/20% split to ensure robust model training and evaluation.
        - Preliminary Model Implementation: Several algorithms, including Decision Trees, Neural Networks, Random Forest, Logistic Regression, and Naive Bayes, are implemented, and trained using the training set for the purpose of comparison.
        - Final Model Selection: After evaluating performance, reliability, and interpretability of each model, select the best fit for this research.
        - Pattern Extraction: The selected model was thoroughly assessed and visualized, leading to the extraction of significant patterns and interpretations.

The research conducted three iterations of the designed process using the three tools, and ensured that the entire process, from data ingestion to pattern discovery, is constant, systematic, efficient, and aligned with the research objectives.

viii. Implementation

The implementation phase is critical as it puts the designed process to the test and offers insights of various models. Different algorithms were trained and tested using IBM SPSS Modeler and/or Python (with scikit-learn and PySpark libraries). The performance of these models was evaluated based on several metrics: accuracy (both for training and test datasets) and AUC (Area Under the Curve) for ROC (Receiver Operating Characteristic), true positive rate, and true negative rate. The detailed statistics of each model are as follows:

IBM SPSS Modeler Results:

- Logistic Regression:
  - Training Accuracy: 82.64%
  - Test Accuracy: 82.59%
  - AUC (Training): 0.909
  - AUC (Test): 0.908
- Decision Tree:
  - Training Accuracy: 85.69%
  - Test Accuracy: 86.19%
  - AUC (Training): 0.916
  - AUC (Test): 0.917
- Neural Network:
  - Overall Accuracy: 82.64%
  - True Positive Rate: 86.3%
  - True Negative Rate: 79.3%

Python scikit-learn Results:

- Decision Tree:
  - Training Accuracy: 90.35%
  - Test Accuracy: 85.27%
- Naïve Bayes:
  - Training Accuracy: 76.53%
  - Test Accuracy: 76.06%
- Neural Network:
  - Training Accuracy: 85.22%
  - Test Accuracy: 82.64%

Python PySpark Results:

- Decision Tree:
  - Training Accuracy: 80.00%
  - Test Accuracy: 80.89%
- Logistic Regression:
  - Training Accuracy: 81.09%
  - Test Accuracy: 82.04%
- Random Forest:
  - Training Accuracy: 78.13%
  - Test Accuracy: 78.64%

After thorough evaluation, the Decision Tree model implemented with Python's scikit-learn was selected as the final model, forming the foundation for all subsequent pattern analysis and interpretations. The choice of the Decision Tree model is substantiated by its exceptional

performance. It not only achieves the highest accuracy but is also inherently intuitive. Such characteristics make it an optimal choice for the research's objective, particularly for predicting customers' income levels for ANZ.

ix. Interpretation of Patterns and Results

Hyperparameter Tuning: To optimize the performance of the Decision Tree model, a hyperparameter tuning process was conducted. The GridSearchCV function from the scikit-learn library facilitated this tuning by systematically searching through a predefined set of hyperparameters, aiming to identify the most effective combination. This grid search covered various criteria, including depths, minimum samples required to split a node, and minimum samples required to be in a leaf node.

The results revealed the optimal hyperparameters for the Decision Tree classifier to be:

- Criterion: Gini
- Max Depth: 12
- Min Samples Leaf: 1
- Min Samples Split: 2

When these hyperparameters were applied, the model achieved:

- Training Accuracy: 84.77%
- Test Accuracy: 83.46%
- Test F1 Score: 84.45%
- Test ROC AUC Score: 0.84

The model exhibits an overall good performance, as evidenced by its metrics. The training accuracy of 84.77% is closely mirrored by the test accuracy of 83.46%, which suggests that the model generalizes well to unseen data and there is not a significant overfitting issue. The test F1 score indicates the model's balanced precision and recall. Furthermore, the ROC AUC score of 0.84 on the test data indicates that the model has a good capability to distinguish between the positive and negative classes. Overall, these metrics collectively suggest that the decision tree model is robust and performs well on both training and test data.

Patterns from the Decision Tree Model

Pattern 1: For individuals not married to a civilian spouse, with a lower education level, and without capital gains, the marital status of being "Never-married" doesn't seem to influence their income; they're predominantly classified as earning less than or equal to $85,000. This pattern suggests that single individuals without capital gains and lower educational levels generally have an income below this threshold.

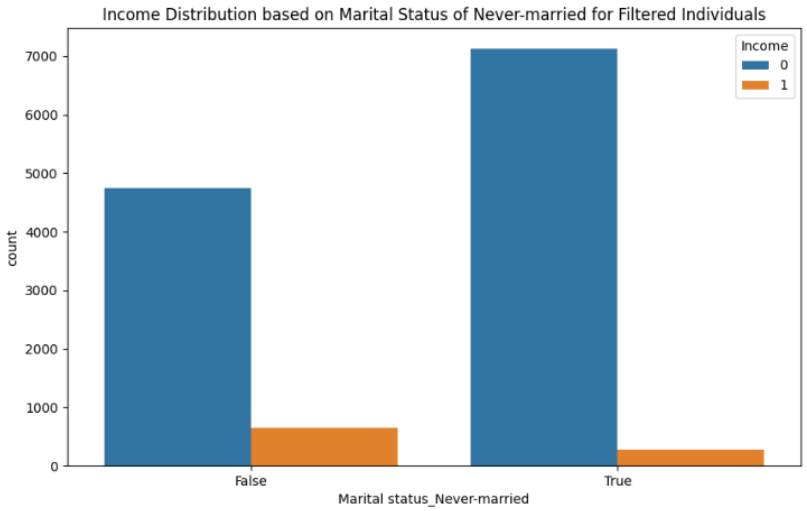


Figure 2: Income Distribution based on Never-married Individuals

Interpretation:

- Description: The decision tree prominently highlights the influence of capital gains in predicting income levels.
- Implication: Capital gains, resulting from asset sales like stocks or real estate, can significantly elevate one's income. Those with higher capital gains likely possess profitable investments or assets, suggesting that investment acumen and asset possession play a vital role in enhancing primary income streams.

Pattern 2: Among individuals not married to a civilian spouse but with a higher education level, the presence of capital gains is a strong indicator of their income being above $85,000, irrespective of their age group. This indicates that young, educated singles with capital gains are more likely to earn above this income threshold.

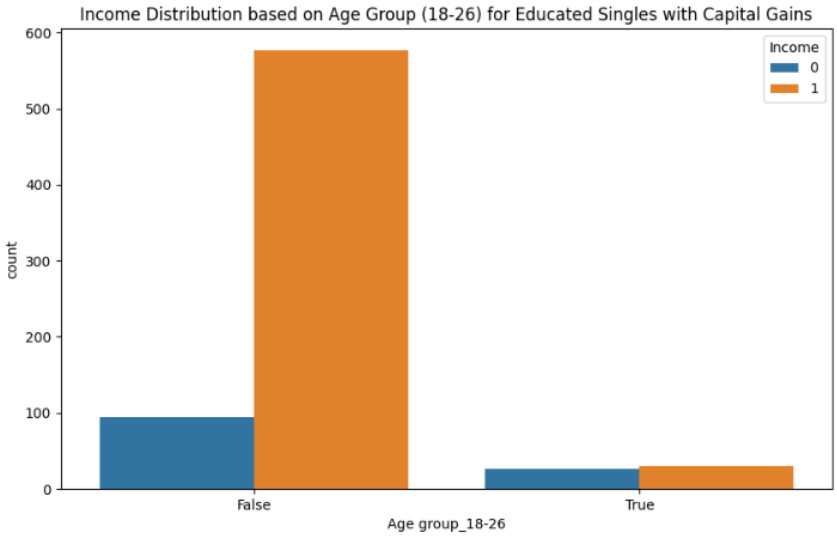


Figure 3: Income Distribution based on Age 18-26 for Highly Educated Singles with Capital Gains

Interpretation:

- Description: The tree emphasizes the Marital status_Married-civ-spouse attribute, denoting the significance of marital status in determining income.
- Implication: Individuals in civilian marriages may benefit from dual incomes or spouse's financial support. Apart from direct financial advantages, marital stability could foster better financial decisions, risk ventures, or earning motivation. However, discerning causation from correlation is essential: those financially stable might marry due to a readiness to provide for a family.

Pattern 3: For those in a civilian marriage and with a higher education level, the absence of capital gains doesn't deter their classification as earning above $85,000. The net capital gain's sign doesn't significantly influence this outcome.

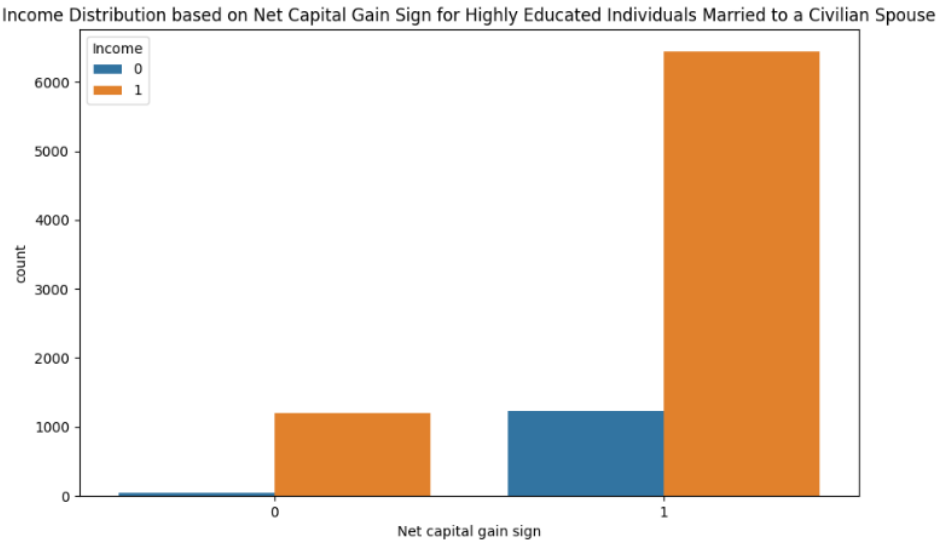


Figure 4: Income Distribution based on Net Capital Gain for Highly Educated Individuals Married to a Civilian Spouse

Interpretation:

- Description: The model often refers to the Education num attribute, indicating the educational level's central role in determining income levels.
- Implications: Higher education often equips individuals with specialized skills, making them more marketable. Renowned educational institutions also proffer a rich network of opportunities, either through alumni, faculty, or on-campus recruitments, facilitating access to higher-paying jobs. This pattern accentuates education's importance beyond skill acquisition, emphasizing the opportunities and networks it offers.

x. Proposed Actions

Applying Knowledge and Deploying the Implementation

Having discovered patterns related to income, marital status, education, and capital gains,

ANZ should focus its marketing strategies on these areas. By understanding the typical income brackets associated with certain demographics and behaviors, ANZ can target its financial products more effectively. For instance:

- Offer investment-related products to individuals with higher education and capital gains, considering their likelihood to earn above $85,000.
- Design unique savings or financial planning products for individuals who are unmarried and have a lower education level, as they typically earn below $85,000.

Monitoring the Implementation

Once the marketing strategies are in place, ANZ should set up a robust analytics dashboard to monitor the responses and effectiveness in real-time. Key metrics might include:

- Customer engagement rate with the new financial products.
- Conversion rates from these targeted marketing campaigns.
- Customer feedback and satisfaction levels.

Maintaining the Implementation

To ensure that the model remains relevant and effective, ANZ should regularly update the dataset and retrain the model. Given the dynamic nature of financial markets and personal income levels, updating the model every quarter or bi-annually could be beneficial.

Enhancing the Solution in the Future

The current model, though effective, can be enhanced further by:

- Incorporating more diverse datasets that capture broader socio-economic factors.
- Exploring ensemble methods to increase prediction accuracy.
- Running periodic feedback sessions with customers to understand any changing preferences or needs, which can then be fed back into the model for refinements.

Recommendations for Research Objectives

1. Marketing Strategy Refinement:
    - Data-Driven Campaigns: Use the findings to target customers with personalized campaigns. For example, reach out to high-earning individuals with investment opportunities and financial planning services.
    - Educational Initiatives: Offer financial education programs to customers with lower education levels to empower them to make informed decisions and potentially elevate their income levels.
2. Strengthening Customer Relations:
    - Personalized Services: Provide specialized financial advisors to high-earning customers to address their unique needs.
    - Engagement Programs: Launch engagement programs for single individuals, focusing on building a strong financial foundation and future planning.
3. Personalized Banking Solutions:

- Custom-Tailored Products: Design banking solutions tailored to the needs of different income brackets, ensuring relevance, and maximizing customer satisfaction.
- Feedback Loop: Regularly collect feedback from customers to refine and tailor banking solutions further.

Further Research and Model Improvement

1. Broader Dataset: Incorporate more granular data points, such as spending habits, to refine the model further.
2. Model Diversity: Explore other machine learning algorithms and techniques to capture diverse patterns within the data.
3. Temporal Analysis: Understand how income levels and associated factors change over time, providing insights into long-term trends and potential future shifts.

Potential Applications in Other Banking Areas

1. Credit Scoring: Use the model to predict potential defaulters by understanding their income and associated factors.
2. Investment Banking: Target high-earning individuals with unique investment opportunities tailored to their predicted income bracket.
3. Asset Management: Understand which customers might have surplus income for investment and offer them tailored asset management solutions.

By addressing the research objectives, ANZ can refine its marketing strategies, strengthen customer relations, and offer personalized banking solutions that are aligned with individual needs, ensuring better customer satisfaction and engagement.

xi. Conclusion

This study aims to better understand ANZ customers' income levels using data analytics. By utilizing tools such as IBM SPSS Modeler, Python scikit-learn, and PySpark, it pointed out the decision tree model from Python scikit-learn as the most suitable for this task.

Key findings included the influential role of capital gains, marital status, and education level on income. These insights can guide ANZ in tailoring its services and marketing strategies to cater more effectively to its clientele. For the future, ANZ can use these insights for personalized banking solutions, enhancing customer relations, and refining marketing strategies. Additionally, this model has potential applications in other banking areas. However, there's room for improvement, and further research can refine the model even more. In essence, by leveraging data analytics, ANZ can make more informed decisions, providing better services to its customers.

References

[1] Trihas, N., Mastorakis, G., Perakakis, E., & Kopanakis, I. (2013). Efficient Customer Relationship Management and Novel e-Marketing Strategies in Tourism and Hospitality. Retrieved from https://dx.doi.org/10.5176/2251-3426_THOR13.09

[2] Ching, R., Cheng, L., Ni, S., & Chen, J. (2007). Applying Data Classification Techniques for Churn Prediction in Retailing.

[3] Osipenko, D. (2016). Credit Card Interest and Transactional Income Prediction with an Account-Level Transition Panel Data Model Using SAS / STAT® and SAS / ETS® Software.

[4] Yoon, C. (2010). Antecedents of customer satisfaction with online banking in China.

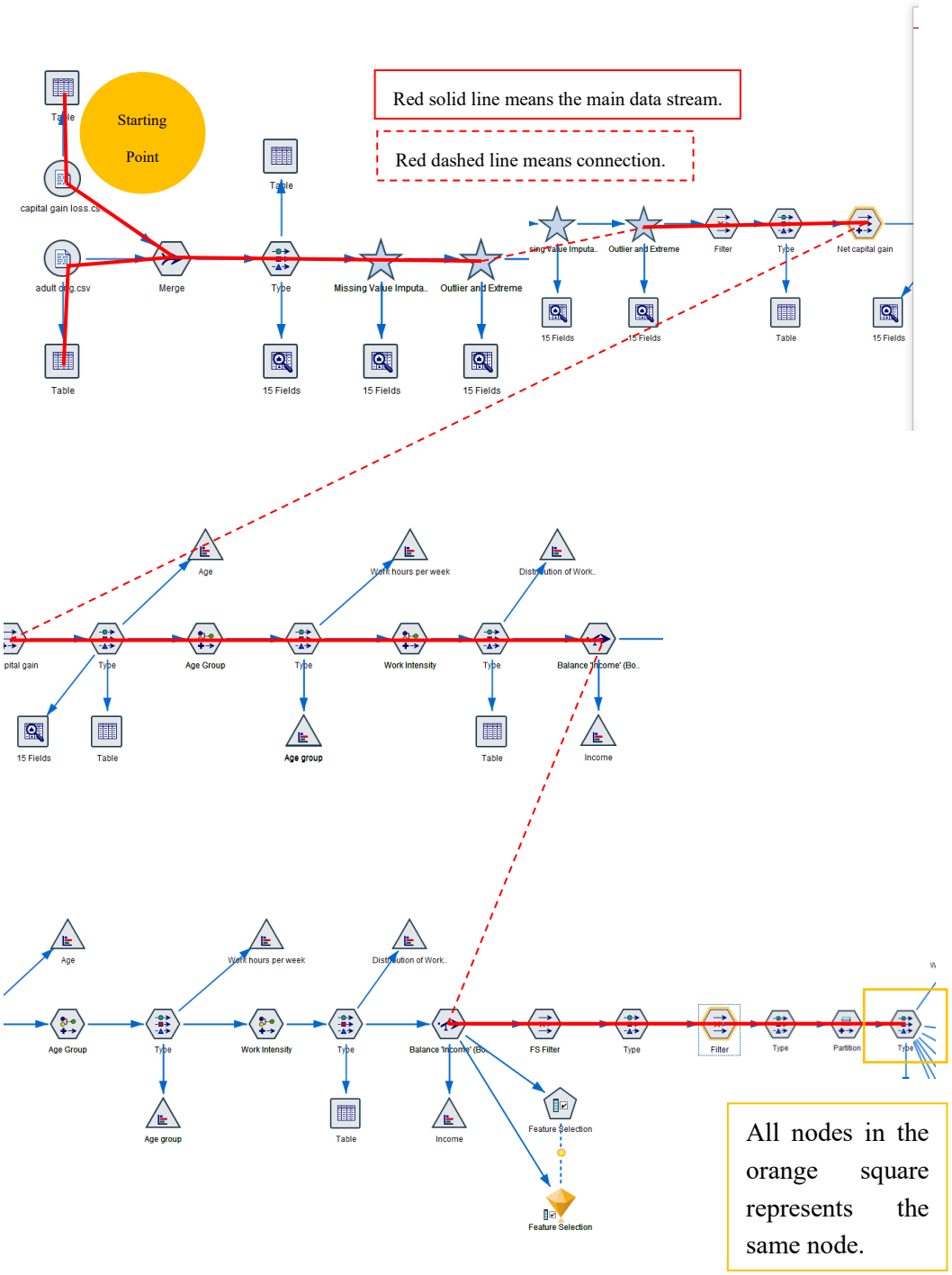
[5] X. Zhang et al., "Personalized Digital Customer Services for Consumer Banking Call Centre using Neural Networks," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 2020, pp. 1-7, doi: 10.1109/IJCNN48605.2020.9206709.

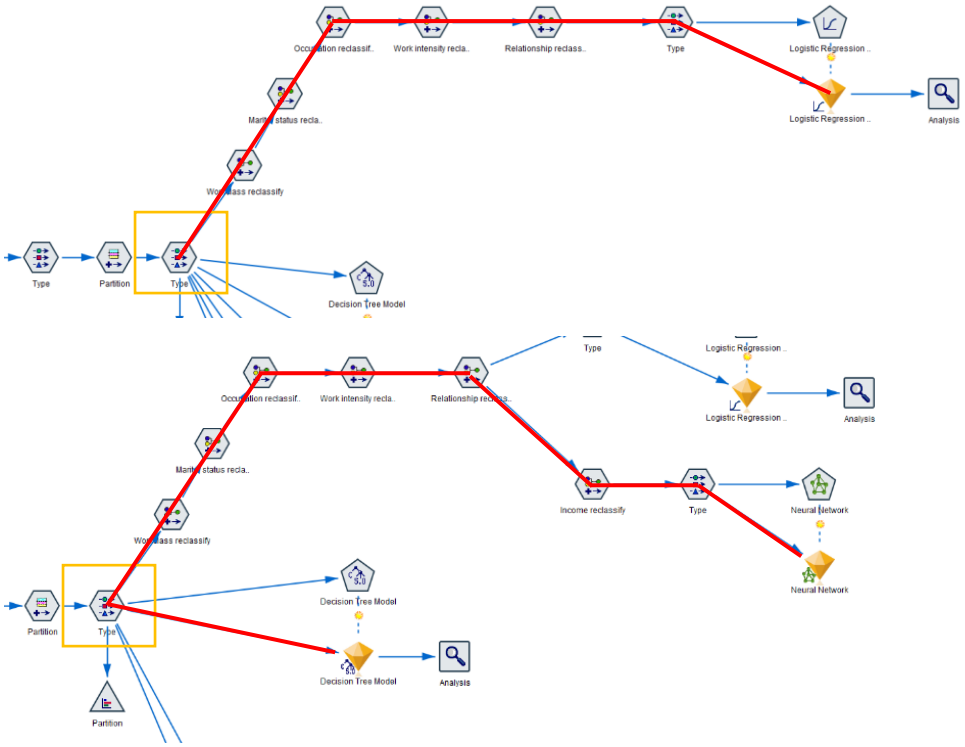
[6] Altwaijri, A.S. (2015). Marketing Strategies and National Culture: an empirical investigation of customers' acceptance of the online banking channel in the context of Saudi national culture.Mogaji, E., Farquhar, J., van Esch, P., Durodié, C., & Perez-Vega, R. (2022). Guest editorial: Artificial intelligence in financial services marketing.

[7] Mogaji, E., Farquhar, J., van Esch, P., Durodié, C., & Perez-Vega, R. (2022). Guest editorial: Artificial intelligence in financial services marketing.

[8] Hassaan, M., Li, G., & Yaseen, A. (2023). Toward an understanding of Pakistani customers' adoption of smart banking services: An extended application of UTAUT2 model with big brother effect and information privacy concern. *International Journal of Bank Marketing*, ahead-of-print(ahead-of-print). https://doi.org/10.1108/IJBM-09-2022-0396

Appendix 1: An Overview of SPSS Modeler Stream




Occupation reclassif..
Work intensity recla..
Relationship reclass..
Type
Logistic Regression ..
Logistic Regression ..
Analysis
Marital status recla..
Work class reclassify
Type
Partition
Type
Decision Tree Model
Type
Logistic Regression ..
Logistic Regression ..
Analysis
Occupation reclassif..
Work intensity recla..
Relationship reclass..
Marital status recla..
Work class reclassify
Income reclassify
Type
Neural Network
Neural Network
Partition
Type
Decision Tree Model
Decision Tree Model
Analysis
Partition

Appendix 2: GitHub Repositories

Implementation using Python Sci-kit Learn: https://github.com/GeorgePiig/722_I3
Implementation using Python PySpark: https://github.com/GeorgePiig/722_I4