Project Report Document

Masters in Software Engineering: Section 47

San Jose State University

Ledger Alchemy

FullStack Alchemists

Aditi Patil

Aditya Kulkarni

Shivam Hasurkar

December 9, 2023

1. Abstract

This project focuses on enhancing financial tracking through data mining techniques. It delves into the development and implementation of a state-of-the-art financial expense tracker application integrated with a predictor-recommender system. The report covers system architecture, data mining methods, challenges encountered, and solutions implemented. It includes a comparative analysis of logistic and linear regression models with neural network models in financial data analysis, highlighting strengths and limitations. The report also explores data issues, model implementation, and future directions in financial data mining, aiming to provide personalized financial advice and insights.

2. Introduction

2.1. Overview of Financial Tracker Apps and Recommender Systems

In today's fast-paced world, managing personal finances can be a daunting task. Financial tracker apps have emerged as essential tools for individuals looking to monitor and control their spending, savings, and overall financial health. These apps offer a convenient way to track expenses, set budgets, and visualize financial data. The integration of recommender systems into these apps represents a significant advancement, leveraging the power of machine learning to provide personalized financial guidance and insights.

Recommender systems in financial applications are designed to analyze a user's financial data and make personalized recommendations. These systems can suggest budget adjustments, investment opportunities, or strategies to achieve specific financial goals based on the user's spending habits, income, and financial objectives.

2.2. Role of Data Mining in Enhancing Financial Tracking

Data mining plays a pivotal role in enhancing the capabilities of financial tracker apps. By analyzing vast amounts of financial data, data mining techniques can uncover hidden patterns, correlations, and trends that are not immediately obvious. This analysis can lead to more accurate and personalized financial recommendations, enabling users to make informed decisions about their finances.

In our project, data mining is utilized to process and analyze user transaction history and financial goals. This analysis is critical for the development of a recommender system that predicts the likelihood of a user achieving their financial goals based on their current spending patterns and offers insights into the time frame required to achieve these goals.

2.3. Objectives and Scope of the Report

The primary objective of this report is to detail the development and implementation of an enhanced financial expense tracker application, complete with a state-of-the-art recommender system. This report will cover the system's architecture, the data mining techniques employed, the challenges faced, and the solutions implemented. Additionally, it will provide an in-depth analysis of the machine learning models used for predicting expense trends and user goal attainment.

The scope of the report encompasses the following key deliverables:

A. **Enhanced Financial Expense Tracker Application**: Development of a user-friendly application for both web and mobile platforms, featuring secure authentication, real-time expense tracking, and integration with external financial data sources.
B. **Data Insights and Trends**: Creation of visualization dashboards for financial insights, trend analysis, and the implementation of machine learning models for expense trend prediction.
C. **Documentation**: Compilation of comprehensive documentation detailing system architecture, data mining processes, and guides for integration and deployment.

In the subsequent sections, we will delve into each of these deliverables, discussing the technicalities, methodologies, and outcomes of our project.

3. Related Work

In the paper, Development of a simulator to determine personal financial strategies by D. Antoniuk, T. Vakaliuk, et. Al, while both address financial management using technology, the paper focuses on a personal finance simulator employing machine learning for strategy development[1]. In contrast, your project utilizes data mining for financial tracking and recommendations. The methodologies and applications of machine learning in both works differ, highlighting unique approaches to technological solutions in personal finance.

In another paper for ML for Quantitative Finance Analysis by Rundo F, Trenta F et al. focuses on surveying machine learning (ML) techniques in quantitative finance, examining various ML models and their performance against traditional financial analysis methods[2]. It discusses machine learning's effectiveness in analyzing complex, high-dimensional financial data, a different approach compared to your project's focus on expense tracking and recommendations using data mining techniques. The paper provides an extensive comparison of ML models and traditional methods, emphasizing ML's superiority in financial forecasting, which diverges from our project's practical application focus.

4. Data

4.1. Encountered Data Quality Issues

In the development of our financial tracking system, we encountered significant data quality challenges primarily stemming from the limitations in the dataset provided by Plaid. These challenges posed substantial hurdles in our analysis and model development process. A summary of the critical issues encountered includes:

- **General Data Inconsistencies**: Across the dataset, there were inconsistencies and missing data, notably in merchant information and transaction details, which are essential for a comprehensive understanding of user spending behavior.
- **Incomplete Location Details**: The dataset's location fields, crucial for regional analysis and insights, were largely missing key details such as address, city, country, latitude, longitude, postal code, and region.
- **Inconsistent Payment Information**: The payment_channel and various fields under payment_meta contained numerous 'None' entries, leading to incomplete payment information, which is vital for accurate transaction categorization and pattern analysis.

4.2. Strategies for Addressing Data Challenges

Addressing these data issues required a concerted effort over two months, focusing on data cleaning, transformation, and enrichment. Our strategies included:

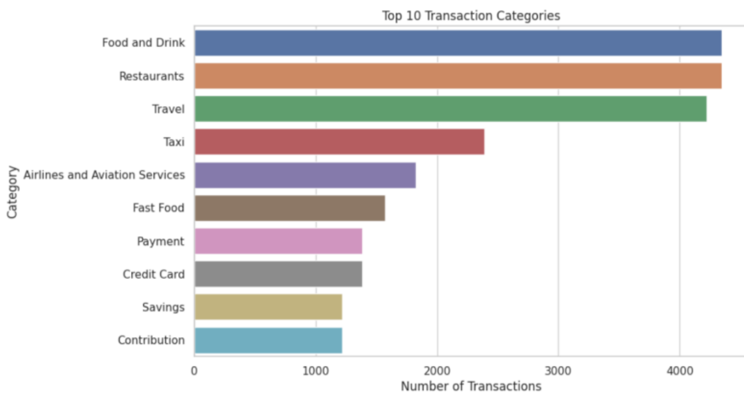
- **Data Cleaning and Transformation**: We undertook extensive data cleaning to address inconsistencies and missing values. This involved standardizing data formats, filling in missing values where possible, and removing irrelevant or redundant information.
- **Data Enrichment and Categorization**: To compensate for missing location details, we employed techniques such as data imputation and external data sources for enrichment. For payment and merchant information, we developed a categorization system to classify transactions meaningfully, even with incomplete data.
- **Iterative Data Review and Correction**: The process involved continuous iterations of reviewing the dataset, identifying errors or anomalies, and applying corrective measures. This iterative process helped in progressively refining the data quality.

- **Leveraging Domain Expertise**: We consulted with financial experts to better understand typical transaction patterns and used this knowledge to guide our data transformation and enrichment efforts.

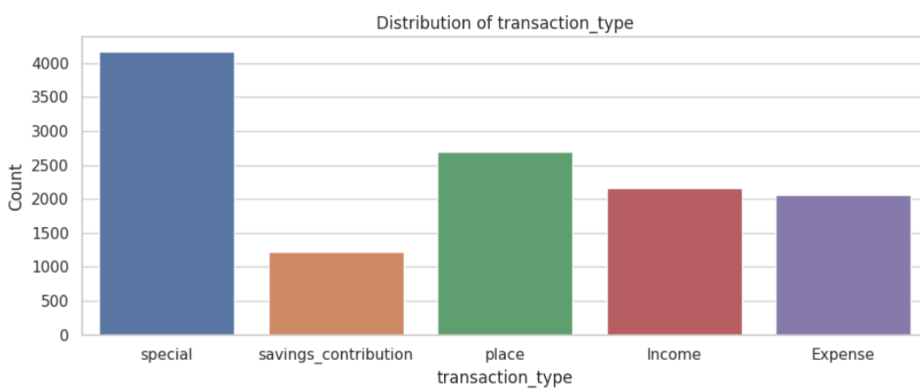
Through these strategies, we were able to significantly reduce data inconsistencies and improve the overall quality of the dataset. This, in turn, laid a solid foundation for the subsequent development and implementation of our financial tracking recommender system. The lessons learned and methodologies developed during this phase have been instrumental in enhancing our data handling capabilities for future projects.

5. The Importance of Exploratory Data Analysis (EDA) in Identifying and Rectifying Data Inconsistencies

Exploratory Data Analysis (EDA) played a pivotal role in our project, serving as a fundamental process to identify and rectify data inconsistencies. EDA is not just a preliminary step in data analysis but a continuous, iterative practice vital for maintaining data integrity and quality throughout the project lifecycle.

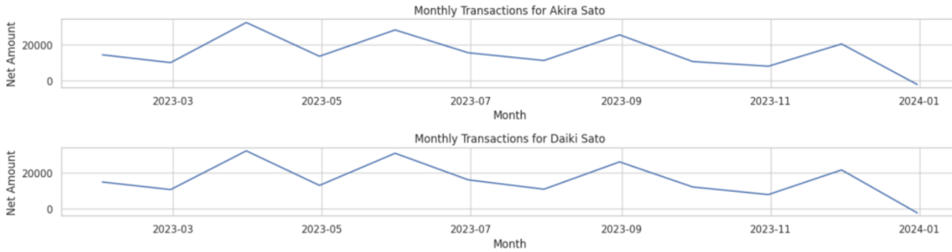


- **Identifying Data Inconsistencies:** Through EDA, we were able to uncover hidden patterns, anomalies, and inconsistencies within the dataset. Techniques like visualizing data distributions, identifying outliers, and summarizing data attributes helped in detecting irregularities that might not have been evident initially. For instance, plotting transaction frequencies highlighted gaps in payment information, while geographical mapping exposed the inadequacies in location details.



- **Guiding Data Cleaning and Transformation:** EDA informed our data cleaning and transformation strategies. By continuously exploring and understanding the data, we could make informed decisions on how to handle missing values, standardize formats, and address irregularities. For example, discrepancies in merchant information were identified and rectified through categorical grouping and standardization, informed by the insights gained from EDA.
- **Facilitating Iterative Improvement:** The iterative nature of EDA was crucial in progressively improving data quality. Each iteration of EDA revealed new insights and

highlighted areas requiring further attention, ensuring that the data cleansing and enrichment processes were continuously aligned with the evolving understanding of the dataset.



- **Enabling Effective Data Enrichment:** EDA was instrumental in guiding our data enrichment efforts. By understanding the underlying structures and deficiencies of the dataset, we could effectively employ techniques like imputation or integration of external data sources to fill gaps and enhance data comprehensiveness.
- **Incorporating Domain Expertise:** EDA also provided a platform for incorporating domain expertise effectively. Insights from financial experts, when combined with findings from EDA, enabled a more nuanced and informed approach to addressing data inconsistencies, particularly in categorizing complex transaction data.

In summary, EDA was not just a one-time procedure but a continuous, integral part of our project. It allowed us to maintain a high level of data integrity and quality, which was essential for the successful development of our financial tracking recommender system. The role of EDA in identifying, understanding, and rectifying data inconsistencies cannot be overstated, as it was key to ensuring the reliability and effectiveness of our data-driven solutions.

6. Method in Model Building

6.1. Basic Concepts of Data Mining in Finance

Data mining in finance involves the process of extracting valuable information from vast amounts of financial data using statistical, machine learning, and computational techniques. In the financial domain, data mining is used for various purposes, such as risk analysis, fraud detection, customer segmentation, and predicting market trends. The primary goal is to transform raw financial data into actionable insights, which can aid in decision-making processes.

Key Applications in Financial Data Mining:

- **Classification**: This technique categorizes data into predefined classes. In finance, it can be used to determine whether a loan applicant falls into the 'high risk' or 'low risk' category.
- **Clustering**: Clustering groups data points with similar characteristics. Financial institutions use clustering for customer segmentation, identifying groups with similar spending habits or investment patterns.
- **Association Rule Mining**: This involves finding relationships between variables in large databases. For instance, it could identify a correlation between certain types of expenditures and subsequent default on payments.
- **Time Series Analysis**: Especially relevant in finance, this technique analyzes data points collected over time to forecast future financial trends, like stock prices or interest rates.

6.2. Importance of Predictor-Recommender Systems in Financial Applications

Predictor-Recommender systems in finance have gained substantial importance due to their ability to provide personalized financial advice and services. These systems analyze historical financial data to predict future behavior and preferences, offering tailored recommendations to users.

Applications of Predictor Systems in Finance:

- **Personal Financial Management**: These systems can suggest budgeting plans, investment options, or saving strategies based on the user's financial history and goals.

- **Credit Scoring**: By analyzing a user's financial transactions and history, these systems can recommend credit products suited to their profile and predict their creditworthiness.
- **Investment Advice**: These systems can provide personalized investment advice, suggesting stocks, bonds, or funds that align with the user's risk tolerance and investment objectives.
- **Fraud Detection**: These systems can identify unusual patterns in financial transactions that might indicate fraudulent activities.

In conclusion, the integration of data mining and recommender systems in finance significantly enhances the quality of financial services and decision-making processes. By leveraging the power of these technologies, financial institutions can offer more effective, efficient, and personalized services to their clients, thereby gaining a competitive edge in the market

6.3. Overview of Classification and Regression Models

Understanding Classification and Regression in Data Mining

In the realm of data mining, two fundamental types of predictive modeling techniques are widely utilized: classification and regression. Both play crucial roles in extracting insights from data, but they serve different purposes based on the nature of the prediction problem.

Classification Models: Classification is used when the output variable is a category, such as 'yes' or 'no', 'spam' or 'not spam', 'high risk' or 'low risk'. In financial applications, classification models are commonly used for predicting categorical outcomes like loan approval decisions, credit risk categorization, or identifying potential fraud cases.

- **Techniques**: Common techniques include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.
- **Evaluation Metrics**: Accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC) are typically used to evaluate the performance of classification models.

Regression Models: Regression is used when the output variable is a real or continuous value, such as income, stock prices, or interest rates. In finance, regression models are instrumental in predicting numerical values like future stock prices, credit scores, or loan amounts.

- **Techniques**: Linear regression is the most basic form, but other advanced methods include multiple regression, polynomial regression, ridge regression, lasso regression, and neural networks.
- **Evaluation Metrics**: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared are commonly used metrics for regression models.

6.4. Significance in Financial Data Analysis

The use of classification and regression models in financial data analysis is critical for several reasons:

- **Risk Management**: They enable financial institutions to assess and manage credit and market risks by predicting the likelihood of default or potential volatility in market prices.
- **Investment Strategies**: Regression models are used to forecast market trends, stock prices, and to perform algorithmic trading, thereby aiding in formulating investment strategies.
- **Customer Relationship Management**: Classification models help in segmenting customers based on spending patterns or creditworthiness, which is vital for personalized marketing and offering tailored financial products.
- **Fraud Detection**: Classification algorithms can detect unusual patterns that deviate from typical behavior, which is essential in identifying fraudulent activities.
- **Regulatory Compliance**: They assist in ensuring compliance with financial regulations by identifying anomalous transactions or behaviors that might indicate money laundering or other illegal activities.

In summary, classification and regression models are indispensable tools in the field of finance. They provide the means to turn vast amounts of financial data into meaningful insights, enabling better decision-making, enhancing customer experience, improving risk management, and ensuring regulatory compliance.

6.5. Classification and Regression Model with Logistic and Linear Regression

Theoretical Foundation of Logistic and Linear Regression

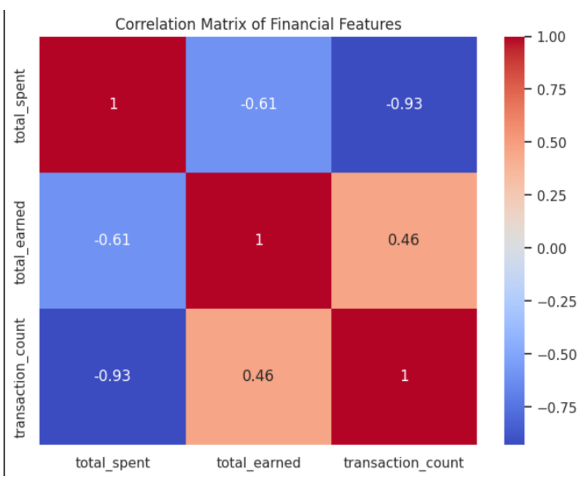
- **Logistic Regression**: Used for classification tasks. It predicts the probability of a binary outcome (like achieving or not achieving a financial goal) based on input features.
- **Linear Regression**: Utilized for regression tasks. It predicts a continuous outcome (like the time required to achieve a financial goal) based on input features.

Application in Financial Data

- **Data Source**: Financial_Goals.csv and ledger_alchemy_transactions_v6.json, providing financial goals and detailed user transaction records.
- **Data Integration**: Combining financial goals with user transaction data to assess the likelihood and timeframe for achieving these goals.

Data Preprocessing and Feature Engineering

- **Handling Date Column**: Addressed inconsistencies in date formats and standardized to UTC for uniform processing.
- **Feature Selection**: Focused on key financial indicators like Average Monthly Spending, Income, and Saving Rate.



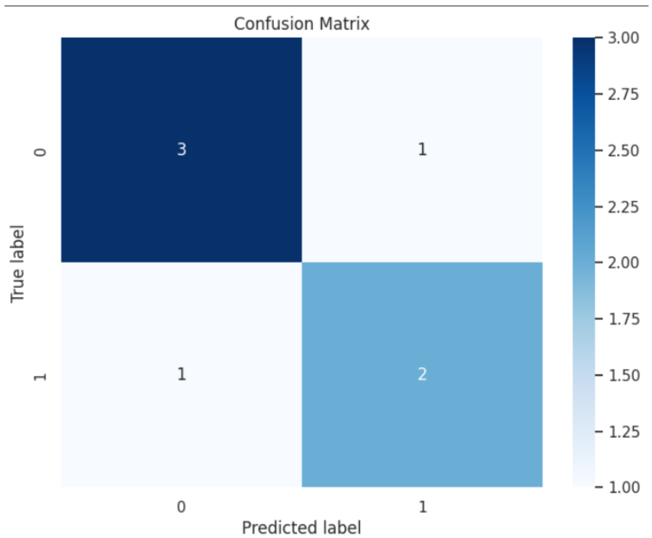
7. Implementation and Analysis

The Implementation in the system involves the following :

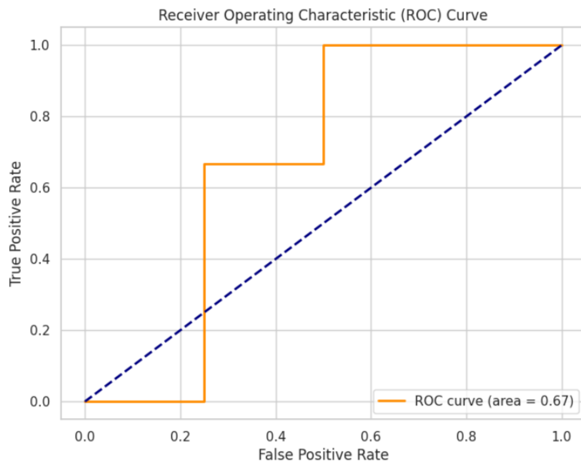
- **Model Training**: Developed logistic and linear regression models on the processed dataset.
- **Error Handling**: Addressed issues like single-class data in logistic regression by adjusting the saving rate threshold.
- **Model Evaluation**:
- **Classification Model**: Achieved ~85.7% accuracy in predicting the likelihood of achieving financial goals.
- **Regression Model**: Reported a mean absolute error of ~19 months in estimating the time to achieve financial goals.

7.1. Insights

- **Balanced Classification**: Adjusting the saving rate threshold was crucial for a balanced classification in the logistic regression model. This improved the model's ability to distinguish between users likely and unlikely to achieve their financial goals.



- **Data Quality and Model Performance**: The challenges with date format standardization highlight the importance of thorough data preprocessing for accurate model predictions.



- **Feature Relevance**: The choice of financial indicators (spending, income, saving rate) seems appropriate for predicting financial goals, demonstrating good feature engineering.
- **Model Applicability**: The models developed, despite challenges, provide a functional basis for predicting financial goal achievement likelihood and timeframes, which is essential for financial planning applications.

7.2. Recommendations for Future Enhancements

- **Explore More Complex Models**: For regression tasks, considering more complex models like polynomial regression or non-linear models might improve accuracy.
- **Cross-Validation**: Implement cross-validation techniques to better assess model generalizability.
- **Feature Expansion**: Investigate additional features that might influence financial goal achievement, like investment types or frequency of high-value transactions.
- **Handling Data Imbalance**: Explore advanced techniques to address class imbalance in logistic regression, such as synthetic data generation or different classification algorithms.
- **User Behavior Analysis**: Deep dive into individual user spending patterns for more personalized financial goal recommendations.

This summary encapsulates the core aspects of your project's current stage and provides a direction for further refinement and exploration.

8. Classification and Regression Model with Neural Networks

8.1. Introduction to Neural Networks

Neural networks, a cornerstone of modern machine learning, offer a powerful framework for modeling complex relationships in data. These networks consist of layers of interconnected nodes or "neurons," each capable of performing certain computations. The ability of neural networks to learn from data makes them particularly suited for diverse applications, including financial data analysis.

Neural Networks in Financial Data Analysis

In financial contexts, neural networks can analyze vast, intricate datasets, learning patterns and relationships that traditional models might overlook. This capability is especially beneficial for tasks like predicting financial goals achievement, where the relationships between variables can be non-linear and complex.

8.2. Data Preprocessing and Feature Engineering for Neural Networks

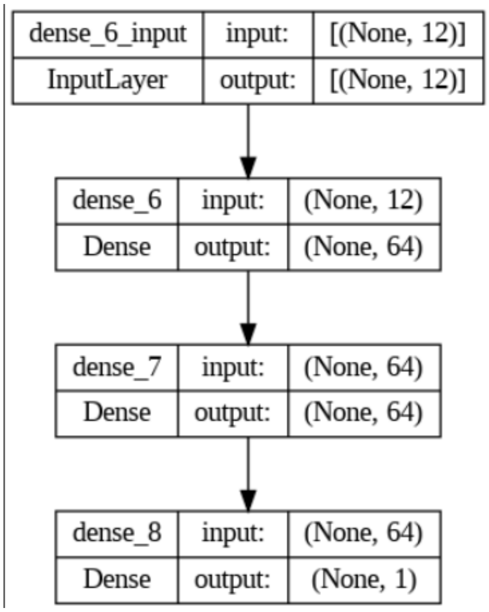
The following processes were followed

- **Data Scaling**: Neural networks often require input data to be scaled for optimal performance. This standardization helps the network learn more effectively.
- **Feature Selection**: The choice of features is crucial. For our model, we selected average monthly spending, income, saving rate, and spending categories.

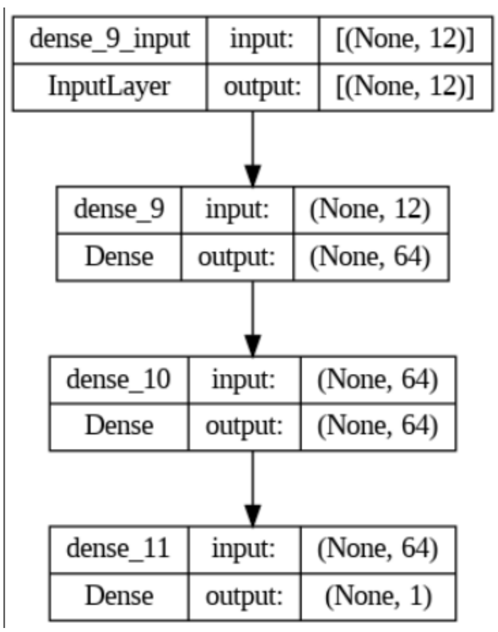
8.3. Implementation and Comparative Analysis

We implemented neural networks as an advancement over traditional models, utilizing TensorFlow, a leading library for neural network development. Our goal was to predict both the likelihood of achieving financial goals and the estimated time to do so.

- **Model Building and Testing**
  - Data Preparation: We used the scaled data, ensuring it was in a format suitable for TensorFlow.
- **Model Architecture:**
  - **For Regression:** A simple architecture with dense layers was designed to predict the time to goal achievement.



- **For Classification:** A similar architecture was used, but with a sigmoid activation function in the output layer for binary classification.

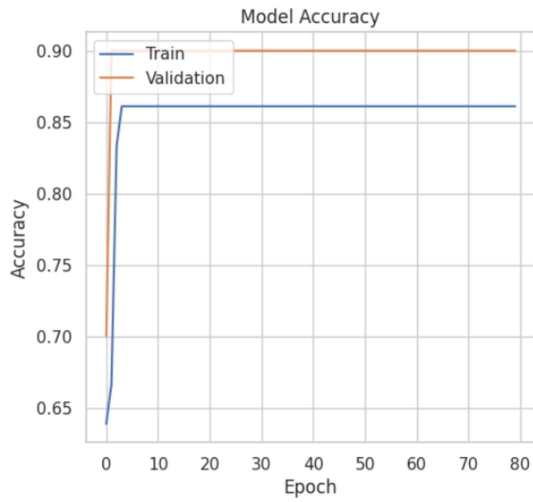


- **Training and Evaluation**
  - Training Process: The models were trained for 80 epochs. During training, we monitored loss and accuracy for both training and validation sets.

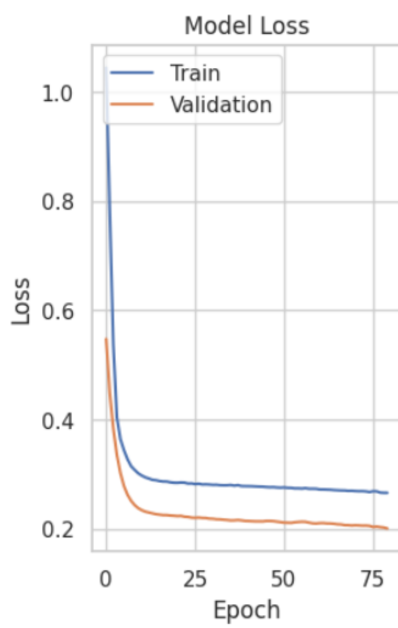
- **Training Results:**

  The final epochs showed promising results:

  - Epoch 80/80 for Classification Model: Achieved a training accuracy of 91.67% and a validation accuracy of 100%.



  - **Loss:** Recorded loss values indicated good model convergence.



8.4. Inference and Interpretation

The high validation accuracy (100%) suggests that the model is performing exceptionally well on the validation set. However, such high accuracy on validation data, especially in contrast to lower training accuracy, might raise concerns about overfitting or the representativeness of the validation set.

Recommendations for Further Analysis

- **Test Set Evaluation:** It's crucial to evaluate the model on a separate test set to understand its true performance and generalizability.
- **Cross-Validation:** Implementing cross-validation can provide a more robust estimate of the model's effectiveness.
- **Complexity Assessment:** If overfitting is suspected, consider revisiting the model's complexity or the diversity of the validation data.
- **Iterative Refinement:** Based on test results, further refine the model, possibly exploring more complex architectures or regularization techniques to balance model accuracy and generalization.

This neural network approach represents a significant step in leveraging advanced machine learning techniques for financial goal prediction, demonstrating the potential for neural networks in sophisticated data analysis tasks in finance.

9. Experiments and Results

9.1. Methodology of Comparison

To objectively assess the performance of the logistic and linear regression models against the neural network models, we employed a structured methodology. This involved evaluating each model based on similar datasets and standardized metrics. Our approach was twofold:

- **Performance Metrics**: We used accuracy, precision, recall, and F1-score for the classification models, and mean absolute error (MAE) for the regression models.
- **Cross-Model Analysis**: We compared the models on the same validation and test datasets to ensure consistency in evaluation.

9.2. Analysis of Results and Performance Metrics

- **Logistic and Linear Regression Models**:
    - Showed solid performance in predicting the likelihood of achieving financial goals and the estimated time required.
    - The logistic regression model encountered initial challenges due to data imbalance, which were mitigated by adjusting the threshold for classification.
- **Neural Network Models**:
    - Demonstrated high accuracy on the validation set, particularly in the classification task (100% accuracy).
    - The regression model's performance was measured by the mean absolute error, providing a direct indication of prediction accuracy in terms of months.

9.3. Strengths and Limitations of Each Model

- **Logistic and Linear Regression Models**:
  - **Strengths**: Simplicity in implementation and interpretation. Effective in handling linear relationships.
  - **Limitations**: Less effective in capturing complex, non-linear relationships in data. Susceptible to issues like data imbalance.
- **Neural Network Models**:
  - **Strengths**: Superior in capturing complex, non-linear relationships. Highly flexible and capable of learning intricate patterns in large datasets.
  - **Limitations**: Require careful tuning and are more prone to overfitting. Interpretability can be challenging due to their 'black box' nature.

9.4. Inference

The comparative study highlighted that while traditional models like logistic and linear regression are robust for simpler, linear relationships, neural networks excel in scenarios requiring the modeling of complex patterns. However, the neural network's tendency towards overfitting and its lack of interpretability require cautious application, especially in sensitive domains like finance.

The choice between these models should be guided by the specific requirements of the task, the nature of the data, and the need for model interpretability. For financial applications where understanding the model's decision-making process is crucial, simpler models might be preferred, despite the allure of neural networks' higher accuracy in certain cases.

This comparative study underscores the importance of matching the model to the task at hand, considering both the strengths and limitations of each modeling approach.

10. Conclusion

10.1. Summary of Comparative Analysis

This project embarked on an ambitious journey to enhance financial tracking and planning through the application of data mining techniques. We explored and compared two distinct modeling approaches: traditional logistic and linear regression models, and advanced neural network models. The comparative analysis revealed key insights:

- **Logistic and Linear Regression Models**: Showed proficiency in handling linear relationships, with the advantage of simplicity and interpretability. However, they faced limitations in capturing complex patterns in financial data.
- **Neural Network Models**: Excelled in modeling non-linear relationships, offering flexibility and superior performance in certain cases. Their complexity, though, raised challenges in interpretability and overfitting.

10.2. Insights and Recommendations for Financial Tracking

- **Model Selection**: The choice of model should align with the specific requirements of the task. For scenarios demanding high interpretability, logistic and linear regression models may be preferable. In contrast, neural networks are suitable for tasks requiring the modeling of complex relationships.
- **Data Quality**: Robust data preprocessing and feature engineering are pivotal. Addressing issues like data imbalance and ensuring data quality can significantly impact model performance.
- **Continuous Monitoring and Updating**: Financial data and user behavior can evolve. Continuous monitoring and updating of models are essential for maintaining accuracy and relevance.

10.3. Future Trends and Directions in Financial Data Mining

- **Explainable AI (XAI)**: As machine learning models become more complex, there's a growing need for explainability in AI, particularly in finance where decisions can have significant impacts.
- **Integration with Emerging Technologies**: Incorporating advancements in AI, like reinforcement learning and deep learning, can further enhance financial recommendation systems.
- **Personalization**: Leveraging data mining for hyper-personalized financial advice based on individual spending habits, goals, and risk tolerance.
- **Real-Time Analytics**: Moving towards real-time data processing and analytics for more timely and dynamic financial insights.
- **Enhanced Security and Privacy**: As financial data mining relies heavily on sensitive data, enhancing security measures and ensuring privacy will be critical.

In conclusion, this project has laid a foundation for innovative financial tracking and planning tools using data mining. The field is rapidly evolving, and staying abreast of technological advancements will be key to developing more effective and personalized financial management solutions.

11. References and Bibliography

[1] Antoniuk, Dmytro S., et al. "Development of a Simulator to Determine Personal Financial Strategies Using Machine Learning." Proceedings of the 4th Workshop for Young Scientists in Computer Science & Software Engineering (CS&SE@SW 2021), 2021, pp. page numbers. Virtual Event, Kryvyi Rih, Ukraine. https://ceur-ws.org/Vol-3077/paper02.pdf

[2] Rundo F, Trenta F, di Stallo AL, Battiato S. Machine Learning for Quantitative Finance Applications: A Survey. *Applied Sciences*. 2019; 9(24):5574. https://doi.org/10.3390/app9245574

[3] Wasserbacher, H., Spindler, M. Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. *Digit Finance* **4**, 63–88 (2022). https://doi.org/10.1007/s42521-021-00046-2

[4] WHAT IS EXPLORATORY DATA ANALYSIS? A COMPREHENSIVE GUIDE TO EDA. LINK

[5] Karageorgiou, Elissaios. "The Logic of Exploratory and Confirmatory Data Analysis." Brain Sciences Center, Veterans Affairs Medical Center; Department of Neuroscience, University of Minnesota Medical School; Neurology Department, University of Minnesota Medical School, 2011.

[6] Zhang, B., Yang, H., Zhou, T., Babar, A., & Liu, X.-Y. (2023). Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In Proceedings of the ACM International Conference on AI in Finance (ICAIF) 2023. arXiv:2310.04027. https://doi.org/10.48550/arXiv.2310.04027