Machine Learning based Expense Analyzer

Abhijeet Bisht
Department of Computer Science and Engineering
Graphic Era University
Dehradun, Uttrakhand, India
abhhiijeet@gmail.com

Deepanshu Upadhyaya
Department of Computer Science and Engineering
Graphic Era University
Dehradun, Uttrakhand, India
upadhyayadeepanshu8@gmail.com

Divyank Singh
Department of Computer Science and Engineering
Graphic Era University
Dehradun, Uttrakhand, India
divyanksingh0702@gmail.com

Randhir Mall
Department of Computer Science and Engineering
Graphic Era University
Dehradun, Uttrakhand, India
Rdm13092000@gmail.com

Dr. Narayan Chaturvedi
Department of Computer Science and Engineering
Graphic Era University
Dehradun, Uttrakhand, India
narayan.cse@geu.ac.in

***Abstract*—The rapid growth of digital transactions and financial data has generated a need for efficient expense management tools. Manual tracking and categorization of expenses can be time-consuming and error-prone, making it essential to develop automated systems that can accurately analyze and categorize expenses. This research paper presents a machine learning (ML) based expense analyzer, designed to automatically categorize, and analyze expenses from various sources, such as bank statements, receipts, and invoices.**

Keywords- Expense Analyzer, Linear Regression, Longest Short-Term Memory (LSTM), K-Nearest Neighbor (KNN).

I. INTRODUCTION

Expense tracking is an essential aspect of personal and business financial management. With the advent of modern technologies, managing expenses has become more convenient than ever. In our day-to-day life, we often find it difficult to manage our expenses as we are more into our life. Bachelors often find it difficult at the end of the month to manage their expenses which leads them with literally no money. So, this application make it easier for a user to manage their expenses based on the data. Machine Learning (ML) is one such technology that has gained significant attention in recent years for its ability to perform complex tasks, including expense analysis. Machine Learning based expense analyzers can provide automated and accurate insights into one's spending patterns, enabling individuals and businesses to make informed decisions about their finances.

Traditional applications that are available for budget expenses management have a persistent problem when they are deployed for the market. In those applications, there are fixed proportions of salary divided as a globalized ratios such as 50% for the needs, 30% for the luxury and 20% for the savings just to say so. But with the difference in age groups, these types do not comply with all the individual, so these applications fail as users and age-defined test subject increases.

Personal budget applications have swept the globe in recent years and will continue to do so. Since more individuals are using their phones to manage their finances, applying for Visas, corporate and personal payments for various items, and the impact of coronavirus has also contributed to an increase in cost management applications.

The end of this exploration paper is to explore the eventuality of Machine literacy, grounded expenditure analyzers in furnishing precious perceptivity into particular and business charges. Machine literacy is applied in wide variety of fields videlicet robotics, virtual assistant (like Google), computer games, pattern recognition, natural language processing, data mining, business vaticination, online cab booking system (e.g. estimating surge price in peak hour by Uber app), product recommendation, share market prediction, medical opinion, online fraud vaticination, advisory, BoTs(chatbots for online client support), E-mail spam filtering, crime vaticination through videotape surveillance system, social media services(face recognition in Facebook). Machine literacy generally deals with those updates can also affect noisy slants, which may beget the error rate to jump around, rather of dwindling sluggishly. The paper will discuss the underlying concepts of Machine Learning, its applications in expense analysis, and the benefits of using an ML-based expense analyzer. Additionally, the paper will provide an overview of the current state of the art in ML-based expense analysis and highlight potential areas for further research. Overall, this paper aims to provide a comprehensive overview of ML-based expense analysis and its potential to revolutionize the way we manage our finances. By understanding the capabilities and limitations of ML-based expense analysers, individuals and businesses can make informed decisions about adopting this technology to optimize their expenses and achieve their financial goals.

Main contributions include:
i. Dataset Generation: Using online forms as there is no readily available dataset on popular publicly available websites like Kaggle, and open-source government docs.
ii. Implementing Neural Network based algorithm: Using Long Short-Term Memory (LSTM) for better accuracy rather than algorithms like Linear Regression, Random Forest Algorithm or K-NN.
iii. Training and testing the Model: On unbiased as well as real-time dataset.
iv. Creating an application: To help user manage their expenses on the basis of prediction made by the model.

II. OBJECTIVES

Following are the objectives to fulfil the requirement for the development of ML Based Expense Analyzer model.

1. Collecting dataset using google form as there is no readily available dataset on popular publicly available websites like Kaggle, and open-source government docs.
2. Implementing Neural Network based algorithm like RNN CNN and LSTM for better accuracy rather than algorithms like Linear Regression, Logistic Regression or K-NN.
3. Training and testing the model on unbiased as well as real-time dataset.
4. Creating an application to help user manage their expenses on the basis of prediction made by the model.

III. RELATED WORK

J. Filliben *et al:* The author included building an application making it easier to manage a user's personal finances. This is split into two halves, accessing historical information in an easy to understand way and using machine learning techniques to predict future financial transactions and the security considerations of storing personal finances information are also considered.

This began with a review of the existing commercial personal finance applications and the current techniques used to forecast time-boxed financial data, such as the value of a stock on the stock market, before detailing the design and implementation of the application.

Mohri Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar: "Foundations of Machine Learning" is a comprehensive resource focusing on the analysis and theory of algorithms in machine learning. It covers a wide range of topics, provides exercises for practical application, and the second edition offers new chapters and additional material, making it a valuable reference for students and practitioners in the field.

V. Singh, L. Freeman, B. Lepri, and A. Pentland: Proposed that human spending behavior is essentially social. This work motivates and grounds the use of mobile phone based social interaction features for classifying spending behavior. Using a data set involving 52 adults (26 couples) living in a community for over a year, we find that social behavior measured via face-to-face interaction, call, and SMS logs, can be used to predict the spending behavior for couples in terms of their propensity to explore diverse businesses, become loyal customers, and overspend. The results show that mobile phone based social interaction patterns can provide more predictive power on spending behavior than often-used personality based features. Obtaining novel insights on spending behavior using social-computing frameworks can be of vital importance to economists, marketing professionals, and policy makers.

Zhang, Shichao, Li, Xuelong, Zong, Ming, Zhu: The paper proposed a kTree method to learn different optimal k values for different test/new samples, by involving a training stage in the kNN classification. The proposed kTree method has a similar running cost but higher classification accuracy, compared with traditional kNN methods, which assign a fixed k value to all test samples. An improvement version of the kTree method is proposed to speed its test stage by extra storing the information of the training samples in the leaf nodes of the kTree. This paper proposes a kTree method to learn different optimal kvalues for different test/new samples, by involving a training stage in the kNN classification. This method has a similar running cost but higher classification accuracy, compared with traditional kNN methods, which assign a fixed kvalue to all test samples.

KIM Zi Won : Proposed to make to make a system providing accuracy and working free for all. The author proposed that maximum accuracy and the service should be free for all. This is a huge problem for all of those who are basically working in any organization, the people who are managing the finances of the organizations they have to do all these things manually. The author is tried to automate this process. Secondly it is very much helpful for those who are doing too many online transactions and those individuals who are in their last year, their financial year are very much confused or very much concerned about how and why they are spending their money. Since this project looks very simple and there is very little user interaction but tricky and interesting machine learning models involved for building the system. They trained on a model for future use. The categories are not so simple for dividing them individually since the uploaded file needs to be converted from PDF format to text format and then to make the data consistent and finally provide the data to the machine learning model. After that machine learning model will be trained in such a way that it will give maximum possible accuracy.

IV. PROPOSED METHODOLOGY

E-commerce reviews reveal the customers' attitudes on the products, which are very helpful for customers to know other people's opinions on interested products. Meanwhile, producers are able to learn the public sentiment on their products being sold in E-commerce platforms. Generally, E-commerce reviews involve many aspects of products, e.g., appearance, quality, price, logistics, and so on. Therefore, sentiment analysis on E-commerce reviews has to cope with those different aspects .The problem with public auction is that the participation of the general public is very limited.

The noble intention of the project is to help individuals and businesses gain better control over their finances and make informed decisions about their expenses. This technology utilizes advanced algorithms and data analysis techniques to automatically categorize and analyze expenses, providing valuable insights and actionable information.

One of the primary goals of an ML-based expense analyzer is to promote financial well-being and improve financial literacy. By providing users with a clear breakdown of their expenses across different categories, such as groceries, utilities, entertainment, and more, it helps individuals understand where their money is going and identify areas where they can potentially save or reduce unnecessary spending. This promotes a greater sense of financial awareness and empowers users to make informed choices about their financial habits.

There are various machine learning algorithms as well which support in in making the predictions.These machine learning models, collect the data, re process it, analyze the data and do prediction. Most of these machine learning models are based on supervised learning algorithm. Some of them are here as follows:

A. K-Nearest Neighbor It is one the supervised learning algorithm which is a non-parametric method used for classification and regression of dataset. In order to select the correct value of K, we need to un the code several time and choose the K which reduces the error while helping the algorithm to predict accurate results.

B. Convolutional Neural Network (CNN) are a type of deep learning model designed specifically for processing grid-like data, such as images or sequential data. CNNs are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers. They are known for their ability to automatically learn and extract hierarchical patterns or features from input data. CNNs have revolutionized image classification, object detection, and computer vision tasks by achieving state-of-the-art performance in various domains.

C. Artificial Neural Network (ANN) are a class of models inspired by the structure and functionality of biological neural networks in the brain. ANNs consist of interconnected artificial neurons organized into layers. Each neuron processes input signals and applies a non-linear activation function to produce an output. ANNs can be shallow or deep, depending on the number of hidden layers. They are widely used for tasks such as pattern recognition, regression, classification, and prediction. Training ANNs typically involves adjusting the connection weights through methods like backpropagation.

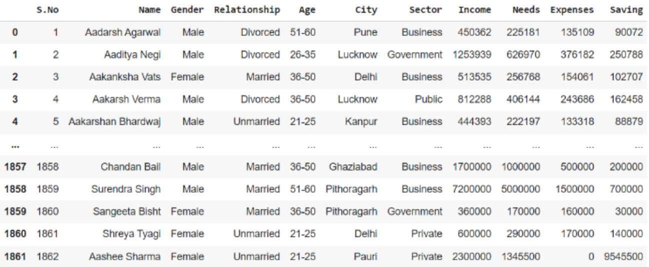
D. Linear Regression is a basic and widely used statistical modeling technique for predicting numerical values based on the relationship between input variables and an output variable. It assumes a linear relationship between the predictors (independent variables) and the target variable (dependent variable). Linear Regression aims to find the best-fit line that minimizes the difference between the predicted values and the actual values. It provides insights into the strength and direction of the relationship between variables and can be extended to multiple linear regression when more than one predictor is involved.

E. Logistic Regression The logistic regression is represented by the sigmoid function graph. It was named after the logistic function. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits .

V. DATA PREPROCESSING

Data pre-processing Data Cleaning and Pre-processing: We have eliminated redundancies from dataset like same name with different spellings or whose names have changed.

Feature extraction From the available data, we can analyses different sectors like public and private by obtaining the statistics from which we can interpret some useful and interesting results. These visualizations can be helpful for the teams and players to understand the areas of improvement and to plan new strategies against opponents.



	S.No	Name	Gender	Relationship	Age	City	Sector	Income	Needs	Expenses	Saving
0	1	Aadarsh Agarwal	Male	Divorced	51-60	Pune	Business	450362	225181	135109	90072
1	2	Aaditya Negi	Male	Divorced	26-35	Lucknow	Government	1253939	626970	376182	250788
2	3	Aakanksha Vats	Female	Married	36-50	Delhi	Business	513535	256768	154061	102707
3	4	Aakarsh Verma	Male	Divorced	36-50	Lucknow	Public	812288	406144	243686	162458
4	5	Aakarshan Bhardwaj	Male	Unmarried	21-25	Kanpur	Business	444393	222197	133318	88879
...	...	...	...	...	...	...	...	...	...	...	...
1857	1858	Chandan Bail	Male	Married	36-50	Ghaziabad	Business	1700000	1000000	500000	200000
1858	1859	Surendra Singh	Male	Married	51-60	Pithoragarh	Business	7200000	5000000	1500000	700000
1859	1860	Sangeeta Bisht	Female	Married	36-50	Pithoragarh	Government	360000	170000	160000	30000
1860	1861	Shreya Tyagi	Female	Unmarried	21-25	Delhi	Private	600000	290000	170000	140000
1861	1862	Aashee Sharma	Female	Unmarried	21-25	Pauri	Private	2300000	1345500	0	9545500

Fig. V.I Dataset

Data Collection - For this project we work on a dataset that contains information about the person and his expenses, for e.g., income, age group, marital status, needs (minimum expense for survival), savings, etc. that is required for making predictions.

The data set for this project is not readily available on popular publicly available datasets like Kaggle, open-source government docs, Google’s Dataset search.

Hence to collect the data we need to manually prepare the huge dataset.

We can use google form to collect as much data as possible, while we can customly scrap different sites to gather as much information as possible to create the data set.

Prediction- It is based on the salary needs luxury and saving columns and based on all the prediction of models. We have used a grouped bar plot to visualize this data.

To acquire the very best version accuracy, the statistics ought to be thoroughly cleaned and pre-processed till it's far nicely perfect. For this we used a Python library along with NumPy, pandas. In order to reap the excellent result for our work, we needed to skip each of our data thru unique gadget gaining knowledge of algorithms along with linear regression, logistic regression, ANN, CNN.

VI.RESULT AND DISCUSSION

This study developed a expense analyzer by collecting data to create our own database using google form as there were no readily available dataset . The reason to collect our own data was to have a unbiased dataset. By training our model using the dataset we were able to predict an average of the expenses of a person belonging to a particular salary range. In order to do this, we used a range of machine learning models, including Linear Regression, Logistic Regression, KNN (K-Nearest Neighbours) and some neural network algorithms such as ANN(Artificial Neural Network), CNN(Convolution Neural Network) and LSTM(Long Short-Term Memory).These models were created using a labelled dataset of information about expenses that included responses from people who had different ranges of salary. We conducted extensive testing and assessment in order to assess the performance of the developed system. A sample of participants completed a google form to get the average expenses of a person belonging to the same salary range. On the basis of the responses given into the trained machine learning models, the algorithm subsequently determined the difference between the expenses of a particular person and the average of the expenses of people belonging to the same salary range.

Classifier	***Accuracy***		
	Needs	***Saving***	***Luxury***
Linear Regressiion	98.67	71.39	94.47
Longest Short Term Memory (LSTM)	99.38	91.02	96.83
Random Forest Algorithm	98.87	87.61	95.56
K-Nearest Neighbors (KNN)	99.12	84.04	94.30

The accuracy of Linear Regression was 83.74%, that of the Random Forest Algorithm was 88.59%, that of the Longest Short Term Memory (LSTM) was 92.86%, and that of the KNN classifier was 91.63%. Accuracy (%).

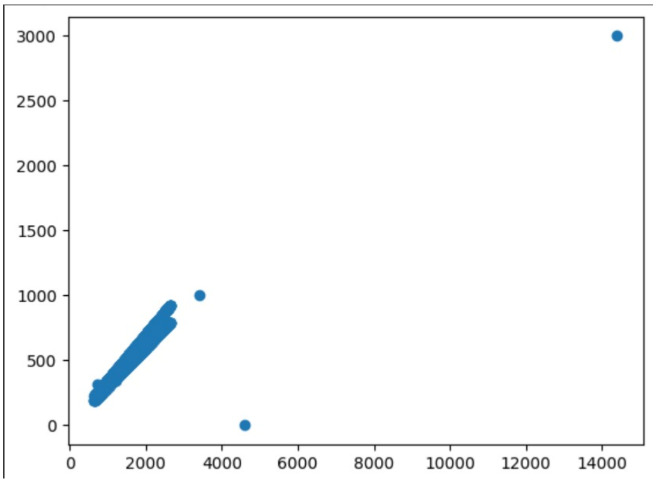


Fig. VI.I Graph plotted for cleaned dataset

VII. CONCLUSION AND FUTURE SCOPE

In conclusion, the ML-based expense analyzer is a powerful tool that offers numerous benefits for individuals and businesses alike. By leveraging machine learning algorithms and techniques, this expense analyzer can effectively automate the process of analyzing and categorizing expenses, saving time and effort for users.

One of the key advantages of using ML in expense analysis is its ability to handle large volumes of data and extract meaningful insights from it. ML algorithms can learn patterns and trends in expense data, allowing for more accurate categorization and identification of spending habits. This helps users gain a comprehensive understanding of their expenses and make informed decisions regarding budgeting and financial planning.

Furthermore, ML-based expense analyzers can provide real-time analysis and reporting, enabling users to track their expenses on an ongoing basis. This timely information allows for better financial control and the ability to identify any potential issues or areas for improvement quickly.

Another significant advantage of ML-based expense analyzers is their adaptability and scalability. These systems can continuously learn and improve over time as they are exposed to more data. As a result, the accuracy and efficiency of expense analysis can increase over time, providing users with increasingly valuable insights. It is worth mentioning that the ML-based expense analyzer is not without its limitations. It relies heavily on the quality and consistency of input data. Any errors or inconsistencies in the data can affect the accuracy of the analysis. Therefore, it is crucial to ensure data integrity and reliability to maximize the effectiveness of the expense analyser. Overall, ML-based expense analysers offer a powerful solution for individuals and businesses looking to gain better control and understanding of their expenses. By leveraging the capabilities of machine learning, these tools can automate the analysis process, provide real-time insights, and support informed financial decision-making.

References

[1] J. Filliben *et al.*, "*Introduction to time series analysis*," in *NIST/SEMTECH Handbook of Statistical Methods*, National Institute of Standards and Technology, 2003.

[2] Shai Shalev-Shwartz, Shai Ben-David, "*Understanding Machine Learning: From Theory to Algorithms*", Cambridge University Press,2014.

[3] V. Singh, L. Freeman, B. Lepri, and A. Pentland, "*Predicting spending behavior using socio-mobile features*,".

[4] KIM Zi Won, "*Comparison between different reinforcement learning algorithms on open AI Gym environment", (Cart-Pole v0)* 2017.

[5] Ankit Choudhary, "*A Hands-On Introduction to Deep Q-Learning using Open AI Gym in Python",* IIT Bombay EEE.

[6] Sherstinsky, Alex, "*Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network*", Physica D: Nonlinear Phenomena,2020.

[7] Zhang, Shichao, Li, Xuelong, Zong, Ming, Zhu, Xiaofeng, Wang, Ruili, *" Efficient kNN Classification with Different Number of Nearest Neighbors"*, IEEE Transactions on Neural Networks and Learning Systems, 2017.

[8] Singh, Himanshu, "*Practical Machine Learning and Image Processing || Basics of Python and Scikit Image*", Volume:10.1007/97,2019.

[9] Mohri Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar, "*Foundations of Machine Learning*",The MIT Press,2018.

[10] Article, IBM, IBM Cloud Education, 17 August 2017
Available: *https://www.ibm.com/cloud/learn/neural-networks*