MACHINE LEARNING MODELS FOR PREDICTIVE ANALYTICS IN PERSONAL FINANCE

Rishabh Kalai[1[0000-0003-0693-128X]], Rajeev Ramesh[1[0000-0003-4889-929X]], Karthik Sundararajan[1[0000-0002-4937-8325]]

[1] Department of CSE, BNM Institute of Technology, Bangalore 560070, India
rishabh.kalai.247@gmail.com

Abstract. Machine learning is an application of artificial intelligence where statistical data is processed by various algorithms that are generally automated in order to produce insights and inferences. It finds many applications in the field of personal finance for portfolio analysis, recommendation engines and even financial forecasting tools. Personal finance management is absolutely crucial to attain financial freedom as well as security. Long term fiscal planning can provide a contingency against uncertainty as well as promote financial stability. The main objective of this paper is to propose RNN based Predictive model for Personal Finance. This paper provides a comprehensive analysis technique that can be utilized to manage the key financial parameters for an individual using machine learning models. In the proposed work, we have implemented three models: A linear regression model for expenditure prediction, RNN based for Stock prediction and Logistic Regression for Retirement Prediction. These models are trained and tested on the basis of both the individual user's data as well as external data pertaining to the economy and the financial markets. In the proposed work, experimental results show an accuracy score of 83.55% for linear regression, 86.7% for RNN and 84.53% for logistic regression, each of which is used for a different phase of the proposed system.

Keywords: Machine Learning, Personal Finance, Linear Regression, Recurrent Neural Networks, Logistic Regression, Retirement Prediction

1 Introduction

Personal finance, as a term, encompasses concepts of management of money, saving and investing. It refers to the entire industry that helps individuals and advises them about financial and investment opportunities [19], budgeting, banking, insurance, mortgages, investments, retirement planning, tax and estate planning and so on. Personal finance management can help an individual effectively plan and reach short-term as well as long-term financial goals progressively throughout their income lifetime and post-retirement as well.

Due to the apparent lack of financial educational courses present in the school curriculum, financial illiteracy among the younger generation as well as adults is prevalent in today's world [11]. Despite adequate resources, there is a widespread behaviorally

anchored mismanagement of money and debt which consequently resulted in inadequate credit scores amongst a large percentage of the general population [21]. The main benefits of personal finance management are long-term financial planning, money management, income and asset protection, investments and retirement/estate planning [4].

With the advent of big data and the rapid proliferation of availability in user related personal data, machine learning has gained favor over the last decade amongst economists and statisticians for data analysis and insight mining [19]. It's plethora of statistical and analytical tools are increasingly being used by businesses to perform comparative analysis studies and even produce personalized recommendations or insights to consumers based on interactions[18]. The purpose of this paper is to propose a model that can be used by an individual to track their personal financial aspects. Algorithms such as linear regression, logistic regression and recurrent neural networks (RNN) are used for the different elements present within the proposed system. In our review we aim to answer the following questions:

RQ1: How is machine learning used in the field of personal finance?

RQ2: Can machine learning improve an individual's finances?

RQ3: Which models would be better suited to address the needs of Personal Finance?

The main contribution of the paper is the detailed description of a plausible system that is capable of tracking, managing and most importantly planning the fundamental features of an individual's finances with the help of machine learning. The proposed system can be utilized to efficiently and systematically improve the finances of an individual and increase financial awareness as well.

2 Review of the Role of Machine Learning in Personal Finance

Machine Learning has a wide array of applications in the finance sector. It is focused on the development of complex algorithms based on mathematical models and data-based model training to conduct predictive analysis whenever new data is supplied.

2.1 Insurance

Insurance is one of the prevalent fields in which machine learning has a lot of use cases [20]. One commonly observed instance of this is where computers run simulation models and compare the pictures of a damaged vehicle provided by an insurance claimant to pictures of a new car in order to assess the extent of damage. The automated system also generally checks and verifies the records of driving history and offers the claimant a quote within a few minutes to settle the claim. This process is more efficient and less cumbersome for the claimant as well as the insurance agency involved compared to the usual tedious procedure [5].

2.2 Chat-Bots/User-Interaction Systems

Machine learning has also been used in chat bots in personal finance management systems [2,10]. It picks up traits and trends from each conversation with a customer and

utilizes it to improve the experience in the future. Natural Language Processing methods are primarily used in this application to identify the user's mood after which, systems can decide if the issue at hand can be solved or requires human intervention and connects the customer to a real life executive [17]. While this may not be directly related to an individual's finances, the use of these systems is prevalent in most of the popular personal finance management applications present in the market today.

2.3 Investment Planning

Another useful application is investment planning and portfolio management. One of the key steps in this process is the determination of the optimal distribution of money between savings accounts and current accounts. Reduction in costs and increase in savings can be attained through the automation of various processes and procedures such as asset management and budgeting [1]. With predictive analysis methods, prediction of stock prices and identification of potential future investments can facilitate the optimization of investments for maximum returns.

3 Proposed Work

Fig. 1 represents the architecture of the proposed system. The proposed work has the following phases: (1) Budgeting and expense Management, (2) Investment portfolio management, (3) Retirement prediction.

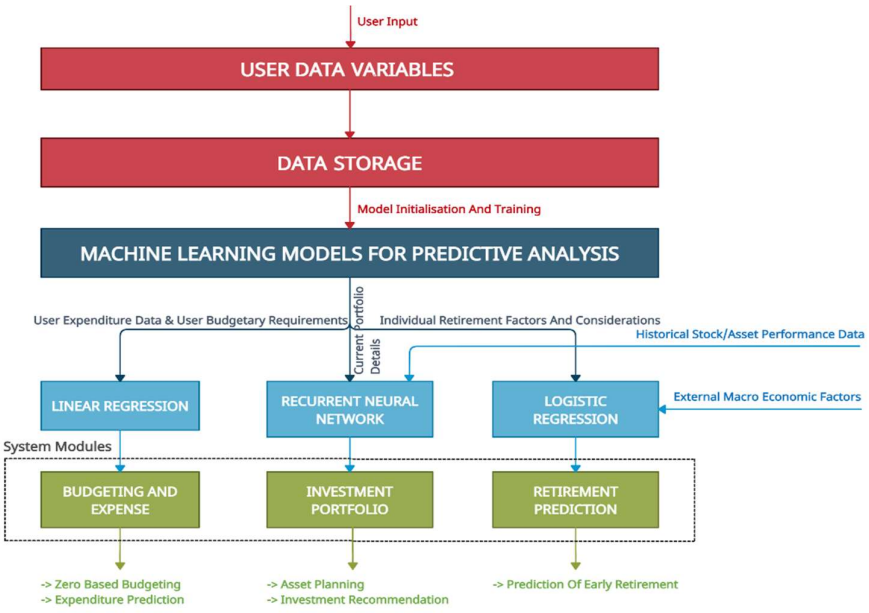

Fig. 1. Architecture of Proposed Model

3.1 Linear Regression Model for Budgeting & Expense Management

Budgeting is the fundamental step in managing an individual's finances [15] and with the advent of automation techniques in finance, budgeting can be managed automatically and more efficiently. The budgeting tool facilitates the tracking of an individual's spending based on the discretionary expenses incurred on a daily basis. The system can then analyze this data that has been gathered to identify spending patterns and areas of potential over-expenditure that the individual might be oblivious to and consequently outline areas of possible savings.

Zero-Based Budgeting.
Zero-based budgeting is a technique based on the allocation of all income and funds to expenses, savings and debt payments. It structures the expenses on the basis of category and also the period in which it occurred. It operates on the principle that total expenditures subtracted from the total income should amount to zero at the end of the month (or the budgeting time period). In the proposed system, we will make use of the zero-based budgeting principle and also predict the expected expenditure for a time period, based on the previous expenses of an individual with the linear regression model. The data recorded for an individual would represent recorded expenses organized categorically with time dimension as shown in Fig. 2.

Date	Category	Amount(Rs.)
2020-01-02	Food	100
2020-01-03	Food	188
2020-01-06	Food	525
2020-01-09	Clothing	228
2020-01-10	Food	100
2020-01-11	Bills	235
2020-01-15	Entertainment	500

Fig. 2. Categorized Expenditure Dataset Used for Linear Regression Model

Linear Regression
Linear regression is a supervised machine learning model that is used to model a mathematical relationship between an independent or explanatory variable and the dependent or response variable by fitting a linear equation between them. The model utilizes a multivariate linear regression equation to describe the relationship [22]. The response variable y is related to the n dependent variables, $x_1, x_2, \dots, x_n$, such that the linear regression equation describing their relationship is of the form:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon \quad (1)$$

where the intercepts $\beta_1, \beta_2, \cdots, \beta_n$ are regression coefficients and ε is the random error [9]. Since the expenses are broken down and classified in terms of different time-periods and categories while also prioritized. The dependent variable y will be the predicted expense for the time period while the independent variables will be denoted by x based

on the different categories and the total expense incurred in each category. The regression coefficients $\beta_1, \beta_2, \cdots, \beta_n$ are used to represent the average functional relationship between variables of interest.

Fitting Using Least-Squares Method.
The next step in the construction of a linear regression model is the fitting of the data to the linear equation. Since the constants β0, β1, β2... βn are unknown, they will have to be estimated. The least-squares estimation principle provides a way of calculating these coefficients effectively. This is done by the minimization of the sum of the squared residuals(distance between the regression line and the data points) from the line described by the linear equation[14]. The method of least squares used to minimize the residual is described by:

$$\sum_{i=1}^{n} r^2 = \min[\sum_{i=1}^{n}(y_i - y_i')^2] \quad (2)$$

where r is the residual, y_i' is the predicted value by the linear regression equation and y_i is the actual value. Now the model is trained and fitted to predict the values of future expenditure based on past expenditure.

Prediction.
The final step in the process of linear regression is to predict the values for the desired time period. This step gives the final total expected expenditure over all categorical expenses for the desired time period in the future for which it was projected. As such, the future expenses for an individual are obtained as a function of past expenditures and spending patterns.

3.2 RNN Based Model for Investment Portfolio Management

Asset Planning.
An asset is a resource that has economic value, owned by an individual or a company that will bring future benefits. Assets are an important aspect when it comes to investing and future planning. Multiple artificial intelligence and machine learning techniques can be used for asset planning and optimization [1].

Artificial Neural Network (ANN) can be defined as a connection of different nodes that are loosely modeled after the neurons in a brain [6,7]. Feed forward artificial neural networks are commonly used for price forecasting. Generally the output produced is affected by size of training set, degree of freedom of network, physical complexity of problem given [8]. With financial assets and global markets becoming increasingly complex, traditional risk models may no longer be sufficient for risk analysis.

Investment Recommendation.
There are many Artificial Intelligence techniques that can be used for stock prediction such as Artificial Neural Networks, ordinary least square regression, elastic nets, LASSO regression, random forest, among which the most applicable being Long Short-

Term Memory (LSTM), Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). Among these techniques, LSTM and RNN use historical data to predict the prices of any particular stock [12]. Hence, LSTM and RNN are used for identifying long term dependencies for future prediction.

Retrieval of Stock Data and Simple Moving Average.
The first step is to read the historical stock market data through the use of a stock market API. The type of data returned from the stock API is of the type time series, a sequence of numbers in chronological order. The stock API shows the following fields : Open, Highest price of the day, Lowest price of the day, Close and Volume. The closing stock price will be used to train the neural network, as we aim to predict the closing price of the stock for investment advice. Simple moving average (SMA) is used to calculate the average value of a selected range of prices, in this case the closing prices of the stock fetched from the Stock API by the number of time periods in that range [23]. The formula for SMA is :

$$SMA = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3}$$

where $x_1, x_2, \ldots, x_n$ are prices of stock at period n, n is the total number of time periods.

Data Preprocessing and Training Neural Network.
The next step is to prepare the training data. The training data is prepared with weekly stock prices and the returned Simple Moving Average. With the window size being 50, that means the closing values of the previous 50 weeks of the stock as training features, and the SMA as the label feature. Then the dataset is split into 2 parts. 70% is used as training set and 30% is used as validation set.

Now that the training dataset is ready, the next step is to create a model for time series prediction. For the model to learn the sequential time series data, recurrent neural network layers are created and LSTM cells are added to the RNN. Along with optimization algorithms for machine learning, Root Mean Square Method is also deployed with the model. With the use of Root Mean Square, the difference between the predicted values and actual values can be determined. This will enable the model to learn by minimizing the error during the training process.

RNN Algorithm.
Input : Time series stock market data with SMA for every 50 consecutive weeks
Output : Predicted Value of Stock
Step 1: set input_layer_neurons, input_layer_shape
set output_layer_neurons, output_layer_shape
define rnn_input_shape, rnn_input_neurons
define rnn_output_shape, rnn_output_nerons
Step 2: fit model using sequential()
add layer.dense(input_layer_neuron, input_layer_shape)
add layer.reshape(rnn_input_shape)

Step 3: lstm_cells -> []
for each index in layer
push lstm_cells
add lstm_cells to rnn using model.add()
Step 4: train model using model.fit()to obtain trained_model
make prediction using trained_model
Step 5: calculate current state -

$$h_t = f(h_{t-1}, x_t) \tag{4}$$

Where h_t is the current state, h_{t-1} is the previous state and x_t is the input state
apply activation function -

$$h_t = tanh(w_{hh}h_{t-1} + w_{xh}x_t) \tag{5}$$

Where w_{hh} is the weight of recurrent neuron, w_{xh} is the weight of the input neuron
derive output by applying –

$$y_t = w_{hy}h_t \tag{6}$$

Where y_t is the output, w_{hy} is the weight at output layer
Step 6: perform model evaluation to obtain accuracy

Validation and Prediction.
Now that the model has been trained, the next phase is to prepare it for prediction of future values. But prior to that the model has to undergo validation. Since the data has been split into two parts, 70% training and 30% validation, the training set will be used for training the model and the validation set will be used for validating the model. For prediction it uses a window size of 50 which is the closing values of the previous 50 consecutive days. Since the training set is incremented daily, the values of the past 50 days are used as input to predict the value for the 51st day.

3.3 Logistic Regression based Retirement Prediction

Retirement planning is an essential part of personal finance. As the average life expectancy continues to rise [13], retirement planning has gradually become a more crucial part of the financial planning process. The proposed system utilizes a method with which a prediction can be made if the user can retire early (before the designated retirement age) based on various factors that are observed to be the most crucial in the determination of retirement age. These factors are gender, disease, education level, marital status, income and employment status. Macroeconomic factors such as unemployment rate and stock market condition also play a role.

Serial Number	State	County	Gender	Marital_Status	Disease	Education	Yearly_Income	Age	Retired_Early
1	Alabama	Autauga	Unmarried	Female	Moderate Health	Grad School	72,759	55	1
2	Alabama	Baldwin	Married	Male	Moderate Health	High School	1,38,452	55	1
3	Alabama	Barbour	Married	Female	Perfect Health	Grad School	1,41,986	61	1
4	Alabama	Bibb	Unmarried	Male	Marginal Health	Grad School	1,25,771	61	1
5	Alabama	Blount	Married	Male	Late Disease	Grad School	1,12,494	57	1

Fig. 3. Retirement Dataset Used for Modelling Logistic Regression Predictor

Logistic Regression.

Logistic regression is a supervised machine learning model used to calculate the probability of a certain class or event existing. It is a predictive analysis method that is used in describing the data and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, interval, ratio level independent variables [16]. Logistic regression estimates a continuous quantity i.e. the probability that an event occurs compared to a certain threshold that allows taking decisions about classification of the data [3]. The mathematical formula of a logistic model is :

$$l = \log_b \frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n} \beta_i \, x_i \tag{7}$$

here, l is the log odds, $x_1, x_2, \dots, x_n$ are the predictors, $\beta_1, \beta_2, \dots, \beta_n$ are the estimated parameters of the model, b is the base of the logarithm function, n is the number of observations and p is the probability that the response variable is 1. The probability of early retirement is given by:

$$p = S_b(\beta_0 + \sum_{i=1}^{n} \beta_i x_i) \tag{8}$$

where $\boldsymbol{S_b}$ is a sigmoid function with base b, p is the probability of early retirement and $x = \{x_1, x_2, \dots, x_n\}$ will be the vector of explanatory values describing the factors affecting retirement age and the values of β will be the parameters to be estimated. The sigmoid S_b function is used to classify whether the individual described by the specific equation retires before or after the age of sixty.

Table 3.1 Factors Affecting the Age of Retirement

Factors	Value / Representation
Gender	Male/Female
Disease	Moderate Health/Perfect Health/Marginal Health/Late Disease/Early Disease
Education Level	High School/Under-graduation(College)/Graduate School
Marital Status	Married/Unmarried
Yearly Income	Average Annual Income Before Retirement
Employment status	Unemployed/Employed
Age	Age Of the Person Currently

Prediction.
The logistic regression model has been trained and validated on the data of working and retired individuals along with their corresponding data pertaining to the retirement factors. The model is then used to predict whether an individual described by a specific set of values for each of the retirement factors as well as current age will retire before or after the age of sixty, thus enabling augmenting or modification of retirement plans.

4 Experimental Results

In the proposed work three machine learning models were trained for their respective datasets and the models were evaluated on the basis of accuracy of prediction. The accuracy of the models was evaluated using the calculated MAPE (Mean Absolute Percentage Error) value.

$$\text{Accuracy}(\%) = \frac{\text{TP+T}}{\text{TP+FP+FN+T}} * 100 \tag{9}$$

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} |(Actual - Predicted)/Actual)| * 100 \tag{10}$$

$$Accuracy(\%) = MAX(0, (100 - MAPE)) \tag{11}$$

where *n* is the number of observations for which the MAPE is calculated.TP, TN, FP and FN present in Eqn. 9 denote True Positive, True Negative, False Positive and False Negative respectively.

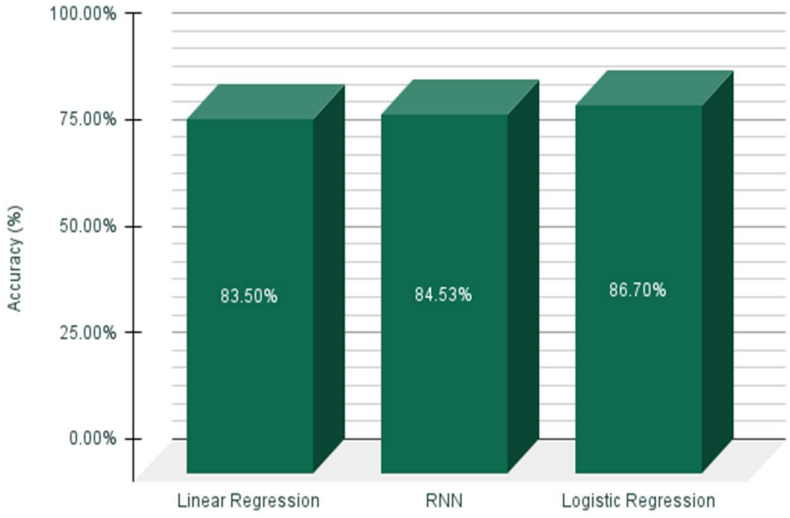

Fig. 4. Performance of the Proposed System

The number of data points for which accuracy is calculated is based on the training/validation dataset size utilized. For the RNN model, the training dataset was set as 80 per cent of the entire dataset, while, the Linear Regression and Logistic Regression models used the entire dataset. Through experimentation, the accuracy of each of the elements of the proposed system were found and is represented in Fig. 4. The accuracy for the linear regression model and RNN was calculated using Eqn. 10 & 11, while, for the logistic regression model, it was calculated using Eqn. 9.

For expenditure prediction, linear regression was utilized in order to generate a trends line that would be capable of being extrapolated in order to forecast future possible expenditure for a given time period. The graph in Fig. 5 displays the linear regression line (trends line) that was generated by regressing the aggregated categorical expenditure over the time period in which the expense occurred. The green line is used to indicate the mean of the entire expenditure dataset which was recorded over the time period for which the linear regression model is trained. This line allows for inference of number of above average expenses incurred. The grey data points are outliers that were imputed.

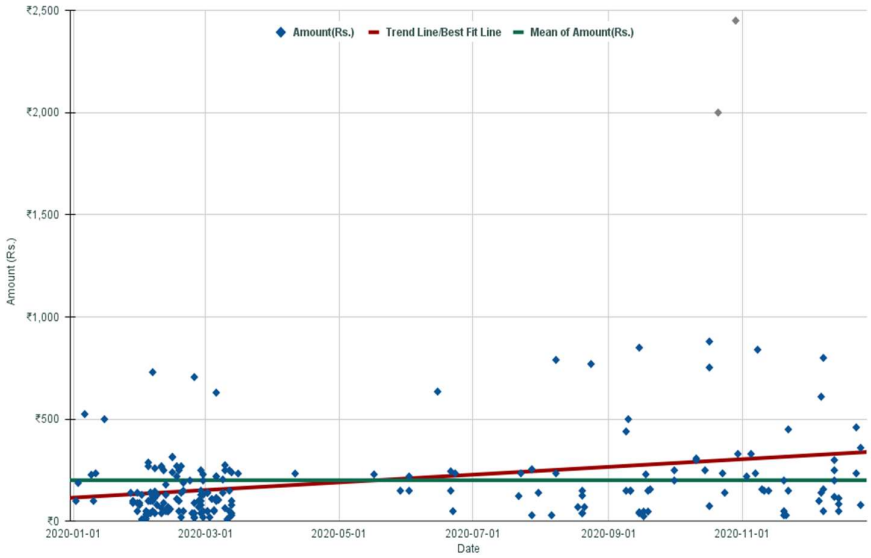

Fig. 5. Linear Regression (Trend Line) of Expenditure vs. Date with Mean of the dataset, used for Expenditure Prediction/Forecasting

Early retirement prediction was conducted on the basis of the macro-factors that are deemed to be the most important determiner/variables. The data was composed of both categorical and numerical variables that were used to perform logistic regression, with the dependent variable being a flag denoting retirement of the individual before 60. The logistic regression predictor modelled was then used to perform predictions of early

retirement based on the testing data. The value 1 is used to indicate true while 0 represents false. The number of correct and false predictions were derived through comparison with actual labels and the confusion matrix shown in Fig. 6 was constructed in order to visualize the number of correct and incorrect classifications. The terms that compose the confusion matrix are: True Positive which denotes correctly classified individuals who retired early, True Negative which denotes incorrectly classified individuals who retired early, False Positive which denotes incorrectly classified individuals who did not retire early and False Negative which denotes correctly classified individuals who did not retire early.

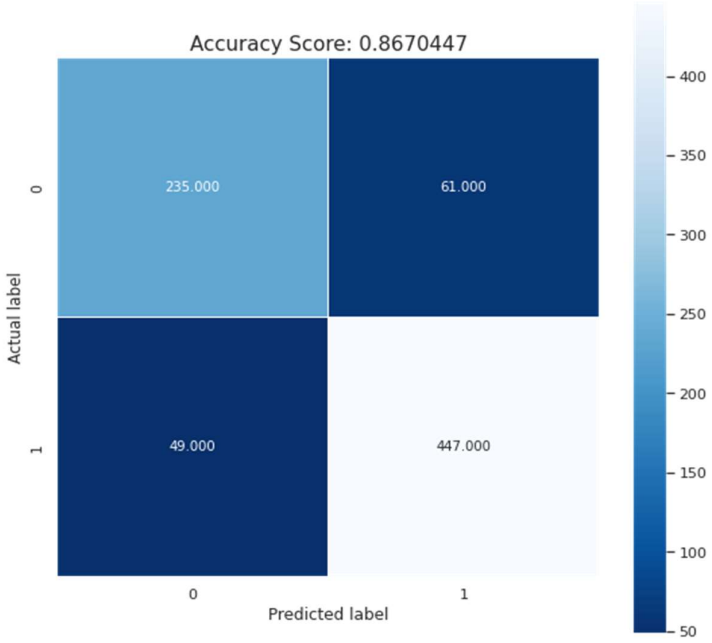

Fig. 6. Confusion Matrix of Modelled Logistic Regression Classifier for Early Retirement Prediction

Fig. 7 and Fig. 8 depict the time series forecasting for stocks: TSLA (Tesla, Inc.) and MSFT (Microsoft Corporation). The blue line depicts the actual price of the stock over the last 20 years, with the green line depicting the predicted value of the stock for the training data by the RNN Model. The yellow trends line is used to show the value of SMA (Simple Moving Average), this is used to ensure that no steep and sudden increase or decrease(outlier) in price of the stock at a short period of time negatively impacts the accuracy of the model significantly. The red line shows the projected value of the stock for the time period into the future for which the prediction is to be made, this is defined by the window size that is set during the retrieval of stock data and computation of

SMA. As such, with a reasonable degree of accuracy, the price of a stock can be forecasted as a function of its historical data.

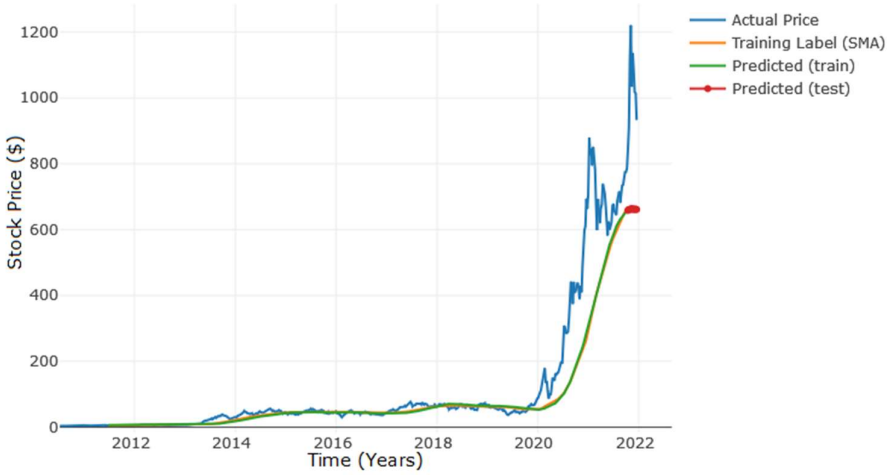

Fig. 7. Time-Series Forecasting for Prediction of TSLA (Tesla, Inc) Stock/Share Price using RNN Model based on SMA values of past performance

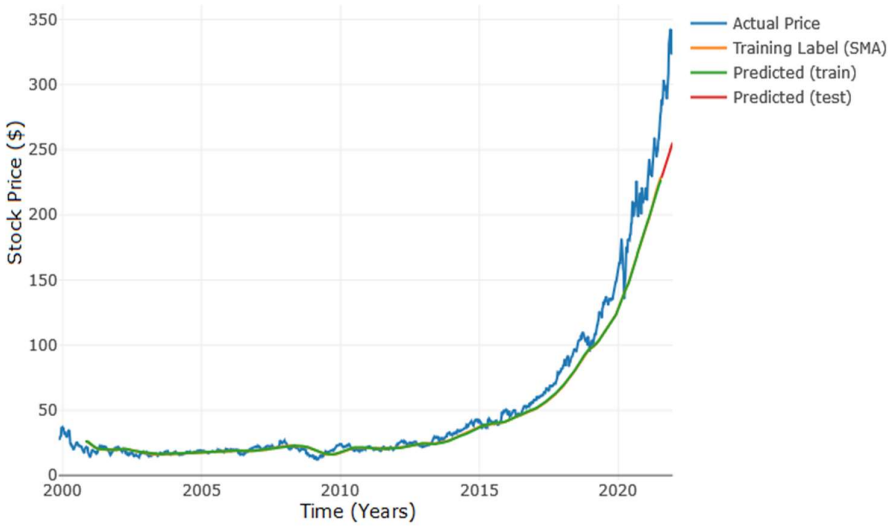

Fig. 8. Time-Series Forecasting for Prediction of MSFT(Microsoft Corporation) Stock/Share Price using RNN Model based on SMA values of past performance

5 Conclusion and Future Scope

Personal finance management is crucial in improving the financial condition of individuals and results in potential increased income and savings amongst a wide slew of other benefits, all of which can ultimately lead to long-term financial security and stability. Machine learning can be utilized in order to further augment and refine the processes present within personal finance management. From the results, it is observed that by using the machine learning algorithms present in the proposed system, predictive analytics can be conducted with a high rate of accuracy for all the three phases of the system. Each of which encompasses a fundamental aspect pertaining to an individual's financial status and condition. Thereby, ultimately enabling a more data-oriented analytical approach for personal finance management. In the future, we would like to explore and identify further applications in the domain of personal finance that can be improved with the utilization of machine learning.

References

1. Bartram, Söhnke M., Jürgen Branke, and Mehrshad Motahari. Artificial Intelligence in Asset Management. No. 14525. CFA Institute Research Foundation, 2020.
2. Dixon, Matthew F., Igor Halperin, and Paul Bilokon. Machine Learning in Finance. Springer International Publishing, 2020.
3. Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5-6 (2002): 352-359.
4. Garman ET, Forgue R. Personal finance. Cengage Learning; 2014 Sep 1.
5. Hanafy, Mohamed, and Ruixing Ming. "Machine learning approaches for auto insurance big data." Risks 9.2 (2021): 42.
6. Haykin, Simon. 2009. Neural Networks and Learning Machines, 3rd ed. New York: Pearson.
7. Aggarwal, Charu C. 2018. Neural Networks and Deep Learning. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-94463-0.
8. Kamruzzaman, Joarder, and Ruhul A. Sarker. "ANN-based forecasting of foreign currency exchange rates." Neural Information Processing-Letters and Reviews 3.2 (2004): 49-58.
9. Kaya Uyanık, Gülden & Güler, Neşe. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences. 106. 234–240. 10.1016/j.sbspro.2013.12.027.
10. Lokman, Abbas Saliimi, and Mohamed Ariff Ameedeen. "Modern chatbot systems: A technical review." Proceedings of the future technologies conference. Springer, Cham, 2018.
11. Lusardi, A. Financial literacy and the need for financial education: evidence and implications. Swiss J Economics Statistics 155, 1 (2019). https://doi.org/10.1186/s41937-019-0027-5.

12. M, Hiransha & Gopalakrishnan, E. A & Menon, Vijay & Kp, Soman. (2018). NSE Stock Market Prediction Using Deep-Learning Models. Procedia Computer Science. 132. 1351-1362. 10.1016/j.procs.2018.05.050.
13. Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2019) “Life Expectancy - OurWorldInData” [Online] Available: https://ourworldindata.org/life-expectancy
14. Mengqiu Kong et al 2019 IOP Conf. Ser.: Earth Environ. Sci. 252 052158
15. Munohsamy, Thulasimani. (2015). Personal Financial Management.
16. Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.
17. Sundararajan, Karthik, and Anandhakumar Palanisamy. "Multi-rule-based ensemble feature selection model for sarcasm type detection in twitter." Computational intelligence and neuroscience 2020 (2020).
18. Sundararajan, Karthik, and Anandhakumar Palanisamy. “Probabilistic Model Based Context Augmented Deep Learning Approach for Sarcasm Detection in Social Media”. International Journal of Advanced Science and Technology, Vol. 29, no. 06, June 2020, pp. 8461-79, http://sersc.org/journals/index.php/IJAST/article/view/25290.
19. Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives, 28 (2): 3-28.
20. “10 Companies Using Machine Learning in Finance to Improve the Entire Industry - Builtin” [Online]. Available : https://builtin.com/artificial-intelligence/machine-learning-finance-examples
21. “Here's how credit scores compare across generations - CNBC.” Available:https://www.cnbc.com/2018/09/25/heres-how-credit-scores-compare-across-generations.html
22. “Linear Regression - Yale” [Online]. Available : http://www.stat.yale.edu/Courses/1997-98/101/linreg.html
23. “Simple Moving Average (SMA) - Investopedia” [Online]. Available :https://www.investopedia.com/terms/s/sma.asp