IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Cash Flow Prediction of a Business Institution Using Deep Learning

Prof. Anagha Chaudhari
Computer Engineering
PCCOE
Pune, India

Aditya Jadhav
Computer Engineering
PCCOE
Pune, India

Uddhav Patil
Computer Engineering
PCCOE
Pune, India

Amitesh Deshmukh
Computer Engineering
PCCOE
Pune, India

Sujit Deore
Computer Engineering
PCCOE
Pune, India

Abstract – **In this paper, we have examined how several deep learning algorithms can be used inthe daily work of a financial organization to estimate cash flow and preprocess data. The challenge is to project the company's various departments' profit and loss so that the firm may prepare for the following year's business expenditures by examining the cash flow prediction. These estimates can then be used to make sure that the right amount of money is kept by businesses in bank accounts for preventing overdraft fees or irrational investment fund usage. The company wants to increase the accuracy of its present manual, ad hoc process for calculating cash flow estimates. Unsupervised neural networks and a variety of other methods like Random Forest, Decision Tree, Linear Regression are being employed, and the results are being used to forecast profit and loss and then the results are being compared at the last to decide which algorithm is best among these algorithms.**

I. INTRODUCTION

In today's world most of the things are moving towards digitalization. One such sector is prediction of profit and loss in business institutions using computers trained by many different algorithms to predict the cash flow pattern. Cash flow prediction is becoming more and more popularas opposed to go through all the stress to observe each expenditure statement on paper and then analyzethem and draw out the conclusion. Analyzing the expenditure on a paper have a huge possibility of human error like taking a wrong value or skipping a value which will lead to draw a wrongconclusion/prediction. Therefore, it is crucial that the conclusion drawn should be correct or else it willlead to false pattern prediction of cash flow. In order to introduce a more efficient and safer way to predict transactions for a business institution we have made and elaborated comparison between NeuralNetworks, Random Forest, Decision Tree, Linear Regression. In simpler terms, these deep learningalgorithms are programs and algorithms that run on a computer and train the prediction model. They are frequently used to automate the prediction of cash flow so that both parties i.e., business institutionand financialist of the company can be confident on the outcomes immediately, without the requirement for a middleman or extra time to pass. They could also automate a procedure so that it only takes action when certain conditions are satisfied.

The major objective of this paper is to determine which algorithm is the best algorithm among the 4 algorithms used and to check if it is possible to create a model that can predict corporate profit and loss more precisely. To accomplish this goal, the company's current and prior year data are collected and employed in a forecasting modelbased on the integration of deep learning and conventional statistical forecasting techniques. We will examine the effects of pre-processing the data into groups and clusters before modelling to highlight the significance of modelling the problem based on homogeneous data sets. As future improvement we are attempting to combine two or more algorithms to improve our model, despite the fact that neuralnetworks and other algorithms have been utilized successfully on a wide range of predicting challenges.

II. MOTIVATION

The major motivation for the paper are the increase in the demand for the cash flow prediction model by business institutions. Many models which are built on different algorithms and are being used in the world today by different businesses are good but not much accurate. Our aim is to build a model which is not only efficient and accurate but also a model which can be used by small businesses and startups. This is a big opportunity in today's growing and demanding world to make our model useful for businesses. The main points of motivation are as follows:

- In today's busy world, where people have busy life schedules, large business working continuously on a daily basis, providing them with efficient cash flow pattern predictionmodel at low price and high efficiency is of utmost important.

- You can determine if your company can pay its debts and make enough money to operate continuously by performing a cash flow study. While persistently positive cash flow is frequently an indication of good things to come, long-term negative cash flow circumstances can signify a future bankruptcy.
- It takes a lot of time to manually review and find patterns in the data for cash flow as the datasets contain lakhs of rows of cash transaction data.
- There is less human intervention when it comes to cash flow prediction flow model, which leads to hassle-free life.

III. LITERATURE REVIEW

In paper [1], Neural Networks: The purpose of this study was to see whether various approaches may help the cash flow forecast of financial firm. It has been shown that the current forecasting system can be improved, however, secret is in the pre-processing stage rather than in the application of advanced forecasting models like neural networks. There are a few problems with this study that need to be highlighted. Theinvestigation's accessible variables were limited, and the results were inconsistent. The study's capacity to recognize and exploit the patterns of certain clients in the modelling would have been enhanced byknowing more about the client. Future studies must look at different data gathering structuresinfluenced by additional knowledge in order to enhance the precision of the modelling approaches used inside the data groups.

In paper [9], Decision Tree: The decision tree methods ID3 C4.5 and CART were used on the dataset.
Compared to other algorithms, decision trees outperform them in terms of accuracy, speed, and precision. The recommendation algorithm is mostly used to find the pertinent information. The research leads to theconclusion that CART is the technique for this dataset that is extremely exact and most accurate amongthe others after a thorough analysis of decision tree algorithms has been carried out.
The use of several datasets for training samples from a substantial data collection has an impact on thetest set's accuracy. Robustness, scalability adaptability, and height optimization problems may be present in decision trees. But unlike other methods of data categorization, decision trees generate a collection of rules that is useful and clear. The most recent research in a number of areas is presented in this article, including the classification of text, smartphone data, disease analysis, and picture and text data. The authors also include descriptions of the datasets they used, the methods/algorithms theyused, and the accuracy outcomes they obtained for decision trees.

In paper [6], Random Forest: The commonly used method of random decision forests was first suggested by Ho in 1995. Breiman's conception of randomization was initially informed by the work of Amit and Geman, who presented the concept of searching across a random subset of the possible node splits in the context of creating a single tree. The design of random forests was influenced by Ho's concept of random subspace selection. The training data must be projected into a randomly selected subspace before each tree or node is fitted in order to add variance among the trees using this method.
This is a popular machine learning algorithm that utilizes the supervised learning approach. It is applicable to ML issues that involve both regression and classification. A prediction model's accuracy is also evaluated using it. It is built on the idea of ensemble learning, which is a method for mixing different classifiers to address difficult issues and enhance model performance.

In paper [4], Linear Regression: A statistical approach used to link variables is regression analysis building a mathematical model to link changes in the dependent variable to independent variables is the main goal of this project. An algebraic solution of the type will often be used to define a regression model. This modelis used to mainly represent relationships between dependent and independent variables. In many research papers researchist use this method for finding dependencies on input and output. It is a supervised learning algorithm. Large datasets can be used to train a regression model.

In financial organizations the methodology to impose regression concepts on a financial mode. Financial institutes use a regression model for predicting sales and budget outcomes. Analysis of Simple Linear regression and Multi Linear regression. Linear equation and slope analysis are used in creating a predictive model.

Terminology related to the Financial sector were involved, the factor of variables especially affectingpayments methods. Which helps in imposing practical model for predicting the profit or loss.

IV. METHODOLOGY

If the forecasting model is properly designed, it will enable more accurate prediction of the amount of money circulation in each unique situation. As long as there is sufficient data to organize the test sample, this procedure can also be automated. For major commercial banks, even a small improvement in predicting accuracy has a discernible economic impact. As a result, redefining current reserves in reserve funds more successfully and lowering the cost of transferring and storing cash are both improved.

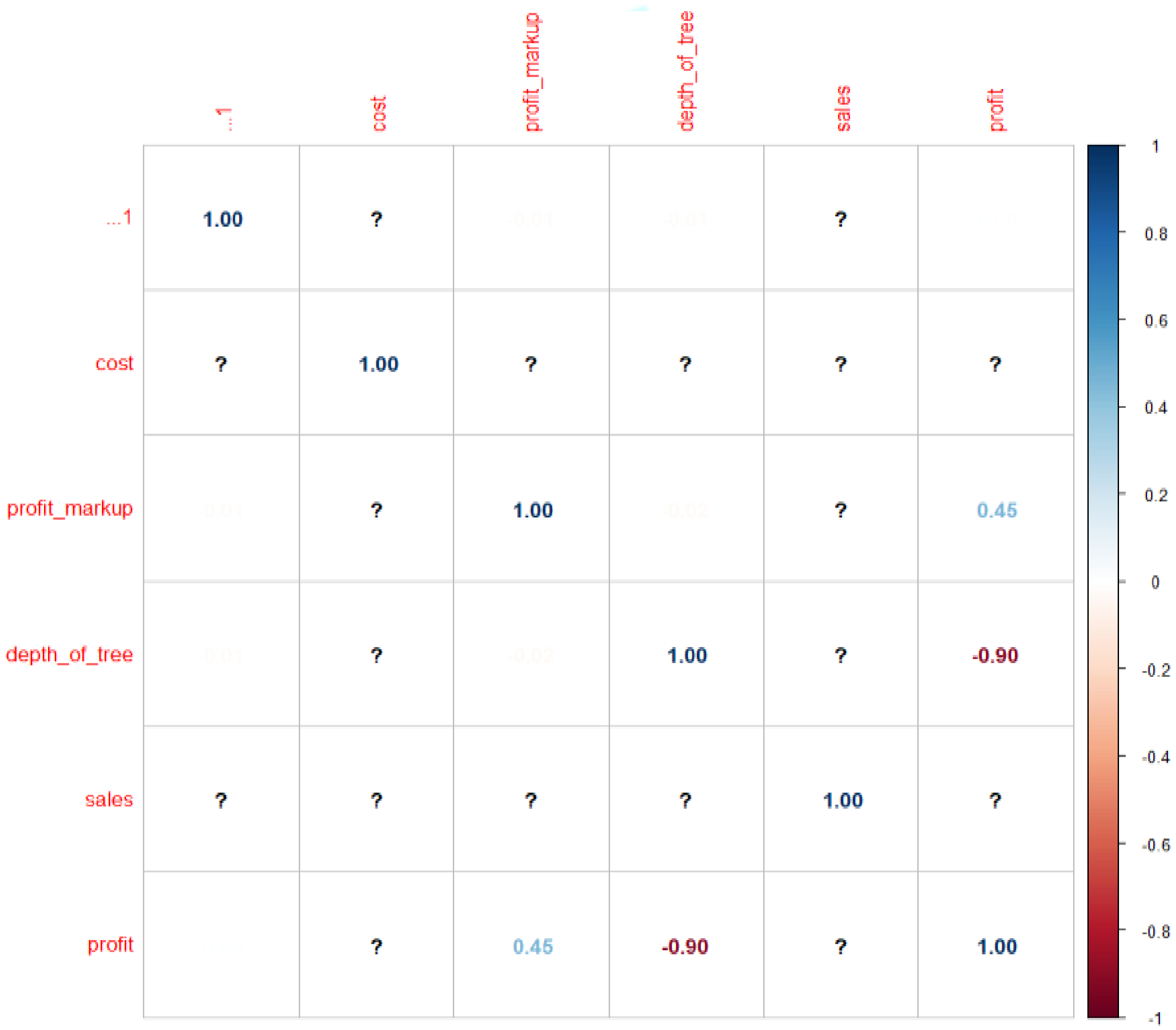

Fig. 1. Correlation Graph

First step in our methodology is to collect dataset of expenditure of businesses that consists the business logic that we are proposing for the system. Then we have to deploy the program on a test network and check its functionalities. The dependencies of the columns present in our dataset are shown in the below correlation graph (fig. 1. Correlation graph).

V. ALGORITHM AND FLOW CHARTS

I. **Neural Network:**

Artificial neurons with synaptic connections between them are the fundamental units of neural networks. ANN actually work as a general-purpose approximator of the function of a number of variables. With a neural network, the analyst's role in the model-building process is quite limited. Specific training procedures can adjust the weighting coefficients to take into account data that has been supplied in advance by the analyst because neural networks can be trained.

The comparison of the model based on MLP with ARIMA and Prophet[7] yielded findings that gave cause for hope that this method of estimating cash flows in our particular situation is promising. The number of input neurons will rise as a result of the first data transformation, which will enhance a
neural network's capacity to predict discrete time series. One of the difficulties with using neural networks is the requirement to produce a training sample that meets the standards for completeness and consistency.

The sample must contain all of the permitted values for the sequence because there are no empty spaces. The consistency of the

sample is defined as the absence of contradictory examples, which may appear in financial time series as a result of the incompleteness of the initial components used in the model.
In such a model, the number of synaptic linkages is determined using the following formula (eq.1):

$$R_w = \sum_{i=0}^{N_l-1} \widehat{N_l} \ \widehat{N_{l-1}} \qquad (1)$$

Formula of the proposed system

II. Linear Regression:

Regression analysis is carried out using independent variables as inputs, regression models a goal prediction value. Finding out how variables and forecasting relate to one another is its main purpose.[2].The type of relation between the independent and dependent variables that each regression model takes into account, as well as the quantity of independent variables utilized, are the two main factors that distinguish them from one another. This analysis estimates the linear equation coefficient by involving single or multiple independent variables which are certainly best for predicting dependent variables values. In pictorial representation as a scattered plot a line is drawn through plots which represent their relationship. Here the equation of the line is linear in the form of (y = mx+c), where m represents the slope of the line, y represents the dependent variable, x represents the independent variable. The line is drawn in a way that minimizes the discrepancies between predicted and actual output values.[3]. This is achieved by using the “least squares” method to discover the best- fit line for a set of paired data. The distance between the point plotted and line is an error in prediction and the actual value, as we represent relation linearly it is difficult to cover each point this increases the error in prediction model and sometimes gives inaccurate results.

In this case to predict cash flow in an organization linear regression is used. The first part of the implementation model is applied to predict profit and loss factors[2.3]. A dataset chosen for training of regression models which have both dependent (overall profit) and independent (budget, debt, liabilities, bills, expenses etc) variables and their values, that are used to define training sets and test sets. By identifying their relationship a linear equation is modified to represent the relationship.[4]

Yi = ß0 + ßxi+ ε (2) Formula for linear regression

- ßxi - the regression coefficient (ß) of the independent variable (x) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)

- $ß_{\circ}$ - the y-intercept (value of y when all other parameters are set to 0)
- Yi - The predicted value of a dependent variable.

MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE).

3. Decision Tree:

Although a decision tree is a supervised learning method, it is frequently chosen for tasks involving classification and regression. It is a tree-structured classifier, where each leaf node represents the classification outcome and inside nodes represent the features of a dataset.

The two nodes in a decision tree are the Decision Node and Leaf Node. Decision nodes are used to produce decisions and have numerous branches, whereas Leaf nodes are the outcomes of decisions and do not have any additional branches.
In paper[24] analyzes the most recent research that have been conducted in a variety of areas, such as the classification of texts, the classifications of user smartphones, the analysis of medical diseases, etc. The capabilities of Decision Tree is to process both numerical and categorical data. simple to understand and interpret. One can picture trees.

The too complex trees that decision-tree learners can produce do not effectively generalizationof input. Overfitting is the term for this. To avoid such problems, mechanisms like pruning, definingthe minimum amount of samples needed at a leaf node, or establishing the maximum depth of the treeare required. Because even minor changes in the data might generate a completely different tree, decision trees could be unstable. The solution to this problem is to use decision trees as part of an ensemble.

Information gain or IG, is a statistical characteristic that evaluates how effectively a certain attribute classifies training instances into the target category. Finding a characteristic that yields the most information gain and the lowest entropy is the key to building a decision tree. Entropy is decreased asa result of information acquisition.

Formula for information gain is:

Information Gain(T,X) = Entropy(T) - Entropy(T,X) (3)

The node with highest information gain is selected as the Root node.

4. Random Forest:

Random forest is a supervised learning method that has labels for our inputs, outputs, and mappings between them. It can be used to solve classification or regression methods. It employs ensemble learning, a technique for resolving complex problems by integrating a lot of classifiers. In paper [22] presents the prevention of financial fraud, for that they used random forest to predict accuracy and their steps are Random Forest is a collection of de-correlated decision trees. Decision trees are used by Random Forest as its fundamental classifier. Multiple decision trees are created by Random Forest for finding accuracy; the randomization comes from two sources: First is a random data selection for bootstrap samples, as in bagging and second is a chance selection of the input features used to create each unique base decision tree. They attempted to improve the random forest classifier's accuracy of prediction in two ways: first, by building numerous decision trees using a random training set and random feature selection, and second, by using the hyperparameters of the random forest classifier in Python. There are two method of random forest from which we can create model - first, Classifier is a method to predict the accuracy of the model using ensemble method. They used decision trees and then taking those trees votings and majority voting will be the answer.

$$\mathrm{H(x)} = \arg\max \sum_{i=1}^{k} I(h_i(x) = Y) \quad (4)$$

H (x) - It is used to combine the classification model.
$\mathrm{H_i}$ - It is used to determine the result of a single decision tree.
Y - It is used to determine the output of the target variable, and I (•).
In the case of cash flow prediction, we employ
the random forest algorithm. We will take data and implement it on that dataset using various techniques, such as data preprocessing. In preprocessing, we use the techniques of normalization and cleaning, and then we split the data into train and test data on which we implement the random forest algorithm to predict the accuracy.

III. COMPARATIVE ANALYSIS

This table gives us brief idea about relative performance of all four algorithms. this helps in concluding that random forest algorithm shows best accuracy, among (Artificial Neural Network, Linear regression, Decision Tree, Random Forest).
After implementing all the four algorithms (Neural Networks, Random Forest, Decision Tree, Linear Regression) on the same dataset the observed results are mentioned in the table (table 1.1):

Table I. Algorithm Accuracy Comparison

Algorithms	**Accuracy**	**Run Time complexity**
Decision Tree	95%	O(k) where k= depth of tree

Artificial Neural Network	95%	O(n-m-h^k – o – i) training samplesm = Features
		k = hidden layersh= neurons
		i = iterations

Linear Regression	88%	O(n)
Random Forest	97%	O(k*m) where k= depth of tree

Thus, from the above table 1.1 we have observed that the algorithm Random Forest shows maximum accuracy when compared with the other algorithms like Decision Tree, ANN, Linear Regression and thus, we can say that random forest is the best algorithm with 97% accuracy.

IV. CONCLUSION

By prolonging the cash flow and escalating the unpredictability and instability of the time needed to recover the money from operational activities, the cash flow analysis trend is thus important for any company. This paper reviews one of the main algorithms (ANN, Random Forest, Decision Tree, Linear Regression) used in today's world in cash flow prediction model.

In this paper we have finally concluded that the random forest is the best algorithm with 97% accuracy among the four algorithms we compared.

REFERENCES

[1] Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNSInternational Joint Conference

[2] https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html

[3] https://link.springer.com/article/10.1007/s42979-021-00592-x

[4] https://iopscience.iop.org/article/10.1088/1757-899X/568/1/012069/meta

[5] https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/

[6] National Research Nuclear University MEPhI, Kashirskoe shosse 31, Moscow 115409, Russian Federation Kazbek Dadtev Boris Shchukin Sergey Nemeshaev
[7] Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet Hans Weytjens, Enrico Lohmann & Martin Kleinsteuber

[8] L. Lokmic and K. A. Smith, "Cash flow forecasting using supervised and unsupervised neural networks," Proceedings of the IEEE-INNS-ENNS InternationalJoint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, pp. 343-347 vol.6, doi:10.1109/IJCNN.2000.859419.

[9] Prediction of Enterprise Free Cash Flow Based on a Backpropagation Neural Network Model of the Improved Genetic Algorithm by Lin Zhu 1, Mingzhu Yan 2 and Luyi Bai
[10] Aleskerov, E, Preisleben, B., Rao, B. Cardwatch: a neural network-based databasemining system for credit card fraud detection.
[11] Nyman R, Ormerod P (2017) Predicting economic recessions using mac

[12] Exploring the financial indicators to improve the pattern recognition of economicdata based on machine learning Xiahui Wei, Wanling Chen

[13] Pattern recognition of financial institutions payment behavior author CarlosLeon, Paolo Barucca, Oscar Acero. 2020

[14] https://learnopencv.com/understanding-feedforward-neural-networks/

[15] L. Lokmic and K. A. Smith, "Cash flow forecasting using supervised and unsupervised neural networks," Proceedings of the IEEE-INNS-ENNS InternationalJoint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, pp. 343-347 vol.6, doi: 10.1109/IJCNN.2000.859419.

[16] Smith, Kate A. "Introduction to neural networks and data mining for businessapplications.".

[17] V. Venugopal and W. Baets, "Neural networks and their applications in marketing management", *Journal of Systems Management*, pp. 16-21.

[18] T. Kohonen, Self-Organisation and Associative Memory, New York:Springer-Verlag.

[19] Amalnik, Mohsen Sadegh. “Cash flow prediction using artificial neural network and GA-EDA optimization.” *Journal of Project Management* (2019): n. pag.

[20] Large group activity security risk assessment and risk early warning based on random forest algorithm

[21] An Application of Ensemble Random Forest Classifier for Detecting Financial Statement Manipulation of Indian Listed Companies: IC3 2018

[22] Sci-Hub | Image Classification of Rice Leaf Diseases Using Random Forest Algorithm | 10.1109/ectidamtncon51128.2021.9425696 (hkvisa.net)

[23] https://www.researchgate.net/figure/Decision-Tree-34_fig1_350386944

[24] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html