Expenditure Predicting using Machine Learning

Vipul, P Vinoth Kumar, Divyansh Dimri, Mayank pathak

***Abstract:** As we know in today's world managing expenses is a very challenging thing. By analyzing our previous expenses, we can predict our upcoming expenses. Now digitalization is everywhere so we can get bank transaction history easily, just by getting the data from transaction history we can predict the estimation of upcoming expense. We can do this using machine learning, machine learning is used in many things one of them is prediction. We are using linear regression algorithm, it is a machine learning algorithm used in prediction. The main aim of this project is to build a system that helps in managing personal finances of the user. This project has mainly three modules, first is to collect the data and prepare it to be used in algorithm, next is to build a network between the algorithm and the dataset. The last one is prediction in which system is going to predict the expenses. Particularly we are predicting the expense of next month. We can also use this system in stock market for predicting the next step if stocks of a company will rise or fall do, this can help us in making money from stock market and manage our expense.*

I. INTRODUCTION

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. And according to Tom Mitchell machine learning is a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. There are mainly two categorized machine learning algorithm which are Supervised learning and Unsupervised. Supervised learning is when an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples comes under the category of Supervised learning. This approach is indeed similar to human learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student then derives general rules from these specific examples. Unsupervised learning is when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own. This type of algorithm tends to restructure the data into something else, such as new features that may represent a class or a new series of un-correlated values. They are quite useful in providing humans with insights into the meaning of data and new useful inputs to supervised machine learning algorithms.

Revised Manuscript Received on November 30, 2019.
* Correspondence Author

Vipul*, Dept. of CSE, SRM University, Chennai, India
P Vinoth Kumar, Asst. Professor (Sr.G) Dept. of CSE, SRM University, Chennai, India
Divyansh Dimri, Dept. of CSE, SRM University, Chennai, India
Mayank pathak, Dept. of CSE, SRM University, Chennai, India

The problem we are doing in this project is based on the supervised learning.

As we are using linear regression to solve this problem. And this algorithm comes under the supervised learning because it's a regression problem and all the regression problems comes under the supervised learning.

Linear regression mainly used in predicting problems. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value based on a given independent variable . So, this regression technique finds out a linear relationship between input and output. Hence, the name is Linear Regression.

The representation of linear regression is a linear equation that combines a specific set of input values x the solution to which is predicted output for that set of input values y. As such, both the input values x and the output value are numeric. The linear equation assigns one scale factor to each input value called a coefficient and it is represented by letter Beta(B). So in simple regression problem, the form of model would be:

$$Y=B0+B1*x$$

Given representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs. By using this we are going to predict the expenses. Here our input is the bank statements of the previous year which we have got from bank. Next step is to give the expected output and after getting both the data, our dataset will be prepared and all we have to do is implement in the algorithm and the further procedure.

II. LITERATURE SURVEY

There are many use of machine learning in today's world and there is one described in this project. These are few existing projects from which the idea was taken and worked over to develop this project that could be of a great use for managing expenses and invest in stock market. The systems are mentioned below:

Z. Huber and H. Perrin developed a project in which the system predict the expenses of next whole year in months, which tells the expenditure of every month of upcoming year.

M. Mulazzani developed a project in which a system will predict if stock market will go up or down.

W. Durand, M. J. Schmidt, R. Dupret, P. Borreli developed a system that will predict the next months expenditure with different categories.

I. Savinkin, Mrs. M, Angel developed a system that helps to decrease the personal expenses by categorizing expenses and pointing out the most expensive category.

R. M. Stair and G. W. Reynolds developed a project in which they made face detection id for the security of mobile by using neural

Network.

III. EXISTING SYSTEM

In the existing system, it will predict the expense of the next month by analyzing the data of previous months. Here the datasets is in months and the system will use the data in algorithm then it will give the predicted value which is the expenditure of the next month. Linear regression is used in this project with the simple linear regression. This project just predict the value on behalf of the dataset which have the small data as this is in month.

Drawbacks of existing systems

- In the previous system the output value is not that accurate because the dataset is not enough as in this project.
- Simple linear regression is used in the project so value of the output can be decline in accuracy.
- The input dataset is not bank statements directly, it is prepared into the list of month wise expenditure.
- This system will not provide the direct connection to bank statements but in present project we just need credit card details.
- It has not any security concerns, the system needs security because it contains bank statements .

IV. SYSTEM SPECIFICATION

Hardware Specification

- Processors: Intel Atom® processor or Intel® Core™ i3 processor
- Disk space: 1 GB
- Operating systems: Windows* 7 or later, macOS, and Linux
- Python* versions: 2.7.X, 3.6.X

Software Specification

- Python 3.5+
- PyCharm Edu 3.5.1
- PIP and NumPy: Installed with PIP, Ubuntu*, Python 3.6.2, NumPy 1.13.1, scikit-learn 0.18.2
- Included Python packages: NumPy, SciPy, scikit-learn*, pandas, Matplotlib

V. PROPOSED SYSTEM

In the proposed system the algorithm will predict the expenditure of next month using one of the supervised learning algorithm of machine learning which is linear regression. There are mainly three steps of this project, first of them is preparing data, Second one is creating the link between the dataset and the algorithm we are using and the last one getting the prediction. We are using machine learning here so we need to get a big data.

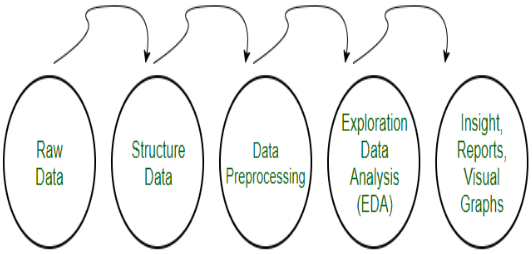
The expenditure predicting system can be categorized into three modules: Data preparation, building network and prediction.

Basic description of modules:

Data preparation: Get the bank statements and prepare the dataset. Prepare a csv file using microsoft excel sheet because we need the dataset in csv format. After that data preprocessing will be performed on the dataset for the missing data.

Building network: In this, we are going to apply the linear regression algorithm. Linear regression is a machine learning algorithm which we are going to apply on the prepared the dataset. So we are building between the algorithm and dataset.

Prediction: In this, the last step is going to take place which is prediction, the algorithm will predict the expenditure of next month by applying the data of the dataset and make prediction. The prediction gets more accurate as the number of execution increases.


Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

A. Data preparation

The process of transforming raw data is the preparation of d ata so that data scientists and analysts can machine learning algorithms to insights and make predictions.The data preparation process can be complicated by issues, It's hard to get every data point in a dataset for every file. Of ten missing data appears as empty cells or a particular chara cter, like a question mark.

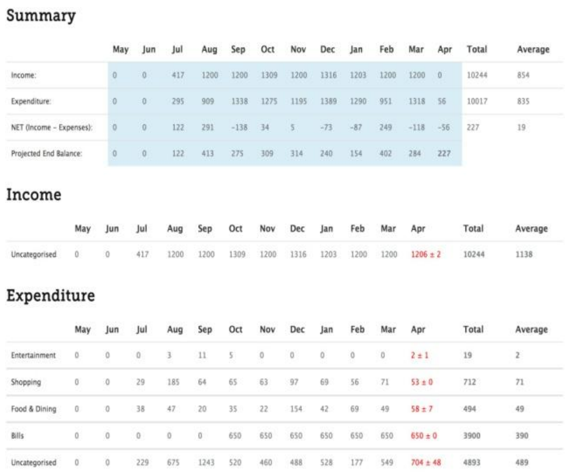
Sometimes it is important to extract data in a different forma t or venue. Consulting domain experts or joining data from o ther sources is a good way to address this.

Even if all relevant data are available, techniques such as fea ture engineering may be needed by the data preparation proc ess to generate additional content that will result in more acc urate, relevant models. Most machine learning algorithms require very specific formatting of data, so data sets usually require some preparation before they can provide useful insights. Many datasets have incomplete, null, or otherwise difficult values for processing an algorithm. If data is missing, it can not be used by the algorithm.

DataRobot automatically performs exploratory data analysis once a user uploads data to the platform, identifying each ty pe of variable and generating descriptive statistics for numer ical records.

It helps analysts and data scientists to analyze data easily, wi thout the need to slice and dice it manually using Excel or ot her similar software.

Data preprocessing is a crucial step in Machine Learning as the quality of the data and the useful information that can be obtained from it directly affects our model's ability to learn, so it is extremely important that we preprocess our data bef ore feeding it into our model.



Summary

	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	Total	Average
Income:	0	0	417	1200	1200	1309	1200	1316	1203	1200	1200	0	10244	854
Expenditure:	0	0	295	909	1338	1275	1195	1389	1290	951	1318	56	10017	835
NET (Income – Expenses):	0	0	122	291	-138	34	5	-73	-87	249	-118	-56	227	19
Projected End Balance:	0	0	122	413	275	309	314	240	154	402	284	227		

Income

	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	Total	Average
Uncategorised	0	0	417	1200	1200	1309	1200	1316	1203	1200	1200	1206 ± 2	10244	1138

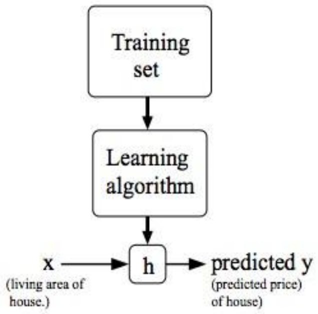
Expenditure

	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	Total	Average
Entertainment	0	0	0	3	11	5	0	0	0	0	0	2 ± 1	19	2
Shopping	0	0	29	185	64	65	63	97	69	56	71	53 ± 0	712	71
Food & Dining	0	0	38	47	20	35	22	154	42	69	49	58 ± 7	494	49
Bills	0	0	0	0	0	650	650	650	650	650	650	650 ± 0	3900	390
Uncategorised	0	0	229	675	1243	520	460	488	528	177	549	704 ± 48	4893	489

B. Building network

Machine learning relies on algorithms to create models that expose trends in data, allowing businesses to discover insights and predict operations to enhance, better understand customers, and solve other business problems. There are many different algorithms, but most scientists are depending on a small set they are familiar with which they are familiar.
Linear Regression is a supervised algorithm for machine lea rning where the predicted output is continuous and has a ste ady slope.
Simple linear regression uses traditional slopeintercept form where m and b are the variables that our algorithm would att empt to "remember" to generate the predictions that are mos t accurate. x is our data input and y is our prediction.

Y=mx+b



Cost function:

The function of prediction is nice, but we don't really need it for our purposes. What we need is a cost function so that w e can begin to optimize our weights. Let's use MSE (L2) as the feature of our value. MSE calculates the average squared difference between the real and expected values of an observation. The performance is a single number representing the value, or rating, of our current weight collection. Our aim is to reduce MSE to improve our model's accuracy. Given our simple linear equation y=mx+b, we can calculate MSE as:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 .$$

Gradient Descent:

The design aims to get the best-fit regression line in linear regression to predict the value of y based on the given input value(x). The algorithm calculates the cost function which tests the Root Mean Squared error between the expected value (pred) and the true value (y) when training the model. The design is aimed at reducing the cost variable. The model needs to have the best value of both x-1 and x-2 to minimize the cost variable. Initially, the model randomly selects x-1 and x-2 values and then iterately updates those values to minimize the cost function until it reaches the minimum. By the time model hits the minimum cost function, it will have the best values for both x-1 and x-2.
Using these finally modified x-1 and x-
2 values in the linear equation hypothesis formula, model pr edicts the value of x in the best possible way.

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Now,

$$\frac{\partial}{\partial \theta} J_\theta = \frac{\partial}{\partial \theta} \frac{1}{2m} \sum_{i=1}^{m} \left[h_\theta(x_i) - y_i \right]^2$$

$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i \right) \cdot \frac{\partial}{\partial \theta_j} \left(\theta x_i - y_i \right)$$

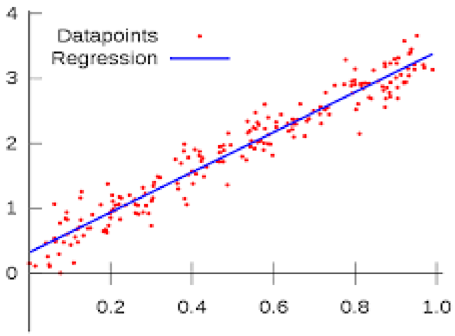
$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} \left[\left(h_\theta(x_i) - y \right) x_i \right]$$

Therefore,

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} \left[\left(h_\theta(x_i) - y_i \right) x_i \right]$$

C. prediction

Prediction ' refers to the performance of an algorithm after it has been trained on a historical dataset and applied to new d ata when predicting the probability of a particular outcome. The algorithm can produce probable values for an unknown parameter for each new data record, allowing the model crea tor to determine what that value will be.Predicting future tra nsactions until they happen is theoretically similar to the wo rk done on the stock market by traders, where the aim is to p redict whether a stock's price will fall or rise to make buy/sel l decisions.
Preifer and Carraway have shown that Markov Chain Model s can be used to model customer relationships with a compa ny and estimate the expected value of an individual custome r marketing commitment.
The combination of the MCM's performance and the weighted average calculator provides the estimate of device expenditure for the class considered. For each sample taken from the MCM, this is measured and the results are passed on to an instance of the Prediction Evaluator to create an average interval of confidence that is eventually displayed on the UI.



The Transaction Markov Chain class takes a list of Transaction objects to construct the Markov Chain Model, for example, and counts the months a transaction takes place, storing it in an associative array. This matrix is read to produce a transition table representing the amount of time a transaction has passed from occurrence to non-occurrenceand visa-versa. The output we got is expenditure of next month which is RS64440 for the given dataset.

VI. RESULT AND DISCUSSION

The results can be summarized as follows:
1)The whole process of understanding the problem domain was undertaken and thoroughly studied various methods for solving the problem.
2)The algorithm worked properly as the prediction it made have a high accuracy, the dataset provided in this problem contains the data of bank statements.
3)As the data in the dataset increases the accuracy of the predicting value will increase, so more amount of data helps in providing more accurate prediction.
4)Finally, the paper had enough content to provide a simple intuition of models and techniques for machine learning to i gnite a tiny spark in the heart of any beginner to try such int eresting ML models later.

VII. CONCLUSION

The project aimed at creating an application that made it eas ier to handle personal finances by automating the typical ste ps people take when making a budget.Existing financial appl ications have been studied, outlining the advantages and disa dvantages of each, as well as defining the features that users found useful by using feedback.. A key piece of the software was addressed, the prediction engine, examining a mixture of proven techniques used for predicting financial expenditure, including MCMs and Weighted Arithmetic Means. The ethical consequences of storing high-risk personal identifiable information are analyzed and understood with respect to safety of application and strong security practices.
This work was then used to evaluate the usability of the soft ware, the design and implementation of the prediction engin e and was used throughout the application to ensure high saf ety standards.Common challenges of design and implementa tion are addressed, outlining the key context and technical k nowledge learned to overcome the challenges faced.
Machine Learning is a required skill nowadays and this paper was to give a basic intuition of some machine learning techniques as well.

REFERENCES

1. V. Singh, L. Freeman, B. Lepri, and A. Pentland, "Predicting spending behavior using socio-mobile features,".
2. KIM Zi Won *Comparison between different reinforcement learning algorithms on open AI Gym environment (Cart-Pole v0)* 2017.
3. 3.Ankit Choudhary *A Hands-On Introduction to Deep Q-Learning using Open AI Gym in Python* IIT Bombay EEE.
4. Andrew NG Machine Learning Course Coursera. Available: *https://www.coursera.org/learn/machine-learning*
5. J. Filliben *et al.*, "Introduction to time series analysis," in *NIST/SEMTECH Handbook of Statistical Methods*, National Institute of Standards and Technology, 2003.

AUTHORS PROFILE

P Vinoth Kumar obtained his Bachelor degree in Computer Science from Vinayaka Missions University, Salem in 2005-2009 and Master degree in the department of Computer Science and Engineering from Vinayak Missions University, Salem in the year 2011. Currently, he is a faculty in the Computer Science and Engineering. Department in SRM Institute of Science and Technology. His main research interest is on Networking and Networking Security.

Vipul is currently pursuing B.Tech in Computer Science and Engineering in SRM Institute of Science and Technology, Chennai, India. Will graduate in 2021. He is very fascinated about technology and frequently writes tech blogs in websites. A tech geek, who loves algorithms and dives deep into the concepts, trying to find efficient ways of solving problems. A machine learning beginner who learned from scratch now is gearing up for future projects and ventures.

Divyansh dimri is currently pursuing B.Tech in Computer Science and Engineering in SRM Institute of Science and Technology, Chennai, India. Will graduate in 2021. He is very fascinated about technology and frequently writes tech blogs in websites. A tech geek, who loves algorithms and dives deep into the concepts, trying to find efficient ways of solving problems. A machine learning beginner who learned from scratch now is gearing up for future projects and ventures.

Mayank pathak is currently pursuing B.Tech in Computer Science and Engineering in SRM Institute of Science and Technology, Chennai, India. Will graduate in 2021. He is very fascinated about technology and frequently writes tech blogs in websites. A tech geek, who loves algorithms and dives deep into the concepts, trying to find efficient ways of solving problems. A machine learning beginner who learned from scratch now is gearing up for future projects and ventures.