

Project and Professionalism

(6CS007)

A2: Project Report

FINANCE

Student Id : 2050089

Student Name : Sajat Prasad Shrestha

Group : L6CG5

Supervisor : Bhanu Aryal

Reader : Rupak Koirala

Cohort : 5

Word count : 2983

Submitted on : 13-11-2021

Contents

[Review on Algorithm 1](#_Toc90588466)

[Regression Model 1](#_Toc90588467)

[Simple Linear Regression 1](#_Toc90588468)

[Multivariate Regression 1](#_Toc90588469)

[Analysis on Regression 2](#_Toc90588470)

[Classification Model 3](#_Toc90588471)

[Support-vector machine 3](#_Toc90588472)

[Naïve Bayes Algorithm 3](#_Toc90588473)

[Analysis on Classification Model 4](#_Toc90588474)

[Research Papers 5](#_Toc90588475)

[Ethereum cryptocurrency price prediction (Poongodi, et al., 2020) 5](#_Toc90588476)

[Similar Systems 7](#_Toc90588477)

[Bibliography 10](#_Toc90588478)

# Review on Algorithm

## Regression Models

### Simple Linear Regression

Linear regression (LR) algorithms are used in order to predict scalar variable for one explanatory variable. (Seltman, 2018) It draws a linear line across two a two-dimensional sample points which includes an independent variable whereas the other one is a dependent variable. The outcome is related to a single regressor. The accuracy of the values is measured by the squared of the constant which will be the residual in the formula y = mx + c. In this formula the y stands for the slope of the linear line that is drew on the graph. This regression model draws a straight line through the graph where there is a Y value for any X value where Y stands for the outcome where as the X stand for the variable which is pre declared. (Lane, n.d.)

### Multivariate Regression

Multivariate Linear Regression algorithms are used to predict scalar variable using multiple correlated variables. (Rencher & Christensen, 2012) Provided a set of such dependent variables along with the predictable variable a model is created that can further be used to predict the value for the unknown data based on probability distribution of the given correlated variables. Such algorithms are best used for prediction, forecasting or error reduction models. (Cohen, et al., 2003)

The model accepts that the association among the dependent variables and the regressor which represents our predictable variable is linear and thus tries to predict the best graphical line that costs the least error. In the graphical representation the Y axis is considered to be a matrix with a series of product of such dependent variables with their own value of theta. Here, theta is considered a constant that is calculated using gradient descent function in order to minimize the cost function. It does so according to its dependency with the X axis. (Izenman, 2013)

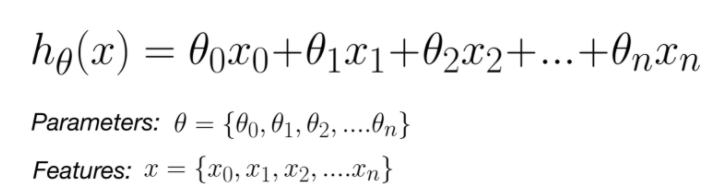


Figure 1: (Nipun Arora, 2019)

### Analysis on Regression

Multivariate Linear regression is the supervised learning algorithm which will be used in order to predict the next expenditure for the user because of its advantages from simple linear regression models. Multivariate models tend to formulate the linear regression line with help of various dependent variables by forming a matrix each with their own constant theta value It will calculate a set of variables in order to calculate the output after figuring a relation between depending variables. The relation between the variables will be calculated using gradient descent whose core objective will be to reduce the RMSE value also known as the cost function. The goal will be to generate a model that will create a personalized best fit regression line for the user in order to predict the next series of expenditure.

## Classification Models

### Support Vector Machine

Support Vector Machines also known as SVM are statistical learning framework used for classifications algorithms. It is a non-probabilistic binary linear classifier. (Vapnik, 1995) It is a supervised machine learning procedure used for not only classification problems but also many regression problems. It works on concepts of hyperplanes in an infinite space where the shortest distance from the training data point is more likely to be categorized as. (Hastie, et al., 2008)

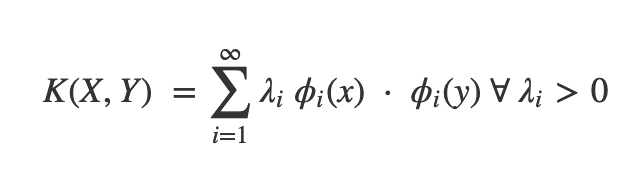
For infinite dimensions support vector machines are formulated as following since the data would be linearly seperable for a hyperplane of infinite dimension . According to the mercer’s theorem, if and only if K(X, Y) is symmetric, positive semi-definite and even continuous, it can be characterized to be such (Shalev-Shwartz, et al., 2016)

Figure (Agarwal, 2019)

### Naïve Bayes Algorithm

Naïve Bayes Algorithm is a supervised probabilistic classifier which works with the Bayes’ theorem to use Bayesian probability concept. (McCallum, 2019) It is generally worked with kernel density estimation in order to accomplish developed precision levels. (Piryonesi & El-Diraby, 2020) This algorithm inherits the method to mark classes to problem occurrences characterized as vectors of feature value. The labels for the classification are drawn from a instantiated set of terms. This algorithm contemplates the features of such classes to be independent to the probability regardless of the probable correlations. This uses the technique of maximum likelihood for parameter estimation.

The conditional probability model is represented by the vector of the probability of n number of independent variables. Here, Bayesian probability to be the product of priority and the likelihood of it happening divided by the evidence. When the features of the vectors are considered to be mutually independent then the formula is the following  
{\displaystyle \varpropto }



Figure (GeeksforGeeks, 2021)

### Analysis on Classification Model

Naive bayes algorithm will be used in order to create a model that classifies natural text into categories by comparing it with a training data set since naïve bayes treats them as independent whereas SVM also focuses on interactions between them. Also, Naïve Bayes algorithms are probabilistic in nature and on the other hand SVM works geometrically. Naïve Bayes algorithm will be trained with a dataset that will contain words that are found in the text that belong to the category where the name of the category will act as a label and so it can probabilistically calculate the category for it.

# Research Papers

## Ethereum cryptocurrency price prediction (Poongodi, et al., 2020)

This research paper emphases upon an analysis based upon two general purposed machine learning models naming Linear Regression and Support Vector Machine using series of inflation and deflation of price in the global market of a blockchain based cryptocurrency called Ethereum with relative to time. Cryptocurrency was a rise of prior failures in digitization of the currency such as Flooz, Beenz and Digicash reasoned to be deficiency of capitals and adaptation. (Tschorsch & Scheuermann, 2016) Bitcoin who brought back this failure also seems to have an issue with imbursement procedure which discusses about two people owning the same coin and was resolved using a dispersed time imprinting service. (Nakamoto, 2008) Cryptocurrency invented by Satoshi Nakamoto keeps ledger of public keys for the sender and the receiver along with the number of coins that were transferred for every transaction that is made by the miners who is rewarded for it. (Poongodi, et al., 2020)

Similar to the analysis a prediction of bitcoin prices was predicted for the closing prices from 2012 through 2018 with SVM using the polynomial kernel function and linear function. This implemented a 10-fold cross validation method where the prediction was made for separate window lengths using different combinations and filter of weight coefficients. (Karasu, et al., 2018) Similarly in this system equation for the linear kernel was used for SVM algorithms which was

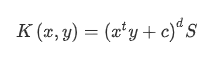


Figure 4:Linear Kernel with SVM (Poongodi, et al., 2020)

And for the radial kernel equation of the algorithm was

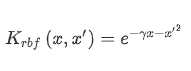


Figure 5:Radial kernel with SVM (Poongodi, et al., 2020)

Where the d variable was taken as lamda equals to the reciprocal of the product of 2 and rho squared which is free and can be optimized. This is also stated to be the basis of the concept of similarity between a pair of samples or observations (Poongodi, et al., 2020)

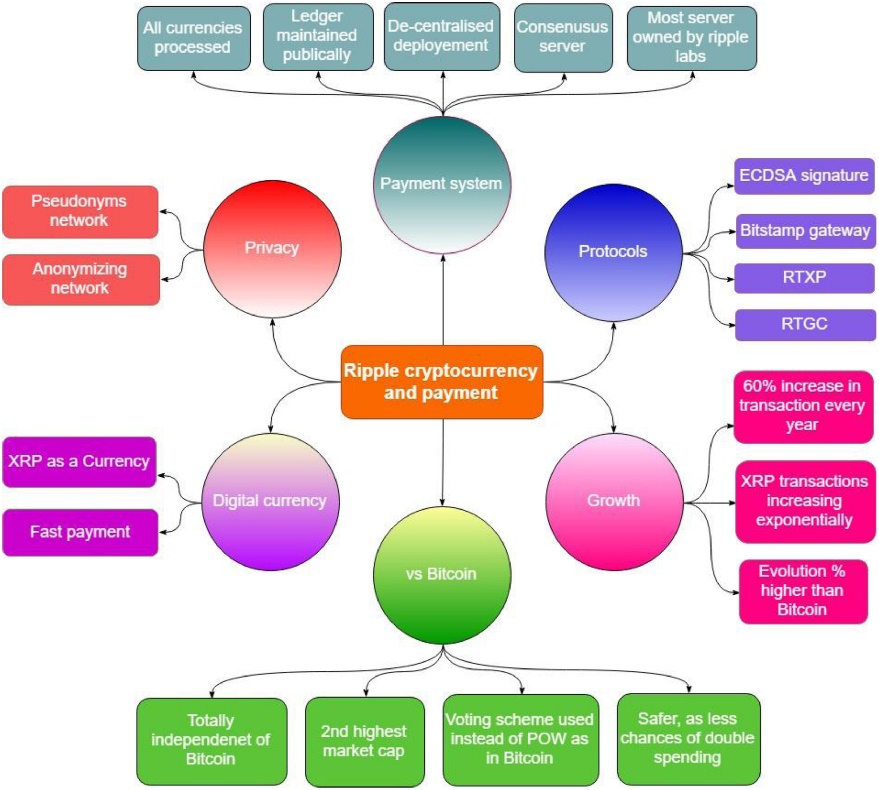


Figure 6: RTXP (Poongodi, et al., 2020)

Similarly, a protocol was used as the RTXP or ripple transaction protocol as given in the figure above. The analysis from the research was that ether saw a growth of 60 percent in transactions every year and also a higher evolution percentage when compared to Bitcoin. In addition, ether was proved better to invest because of the bitcoin’s possibility for double spending. Research concluded provided with enough data that the SVM method had higher accuracy than that of the LR method forming 96.06% to 85.46% respectively. In addition, the model which got proposed was further enhanced with SVM method in order to achieve the final accuracy of 99% which is huge for the variation of uncertainty in prediction of a cryptocurrency. (Poongodi, et al., 2020)

# Similar Systems

Statistical and probabilistic analysis in finance is not a new concept as it has its own interdisciplinary research field known as Econophysics. Use of such study aims to replace the usual prediction methodologies with a more reliable processing of data. It tries to comprehend economic singularities with scientific tools from statistical physics. (shalizi, 2004)

Robert Irons said in his book, The Fundamental Principles of Finance that the standing of accurate estimates cannot be overestimated, valuing possessions too exceedingly will lead to expenses in possessions whose costs are excessive than their returns while undervaluing possessions will lead to lost prospects for growth. (Irons, 2019)

There are plenty of existing systems such as quicken and mint who foresaw this issue. Among the systems mint is the most popular system.

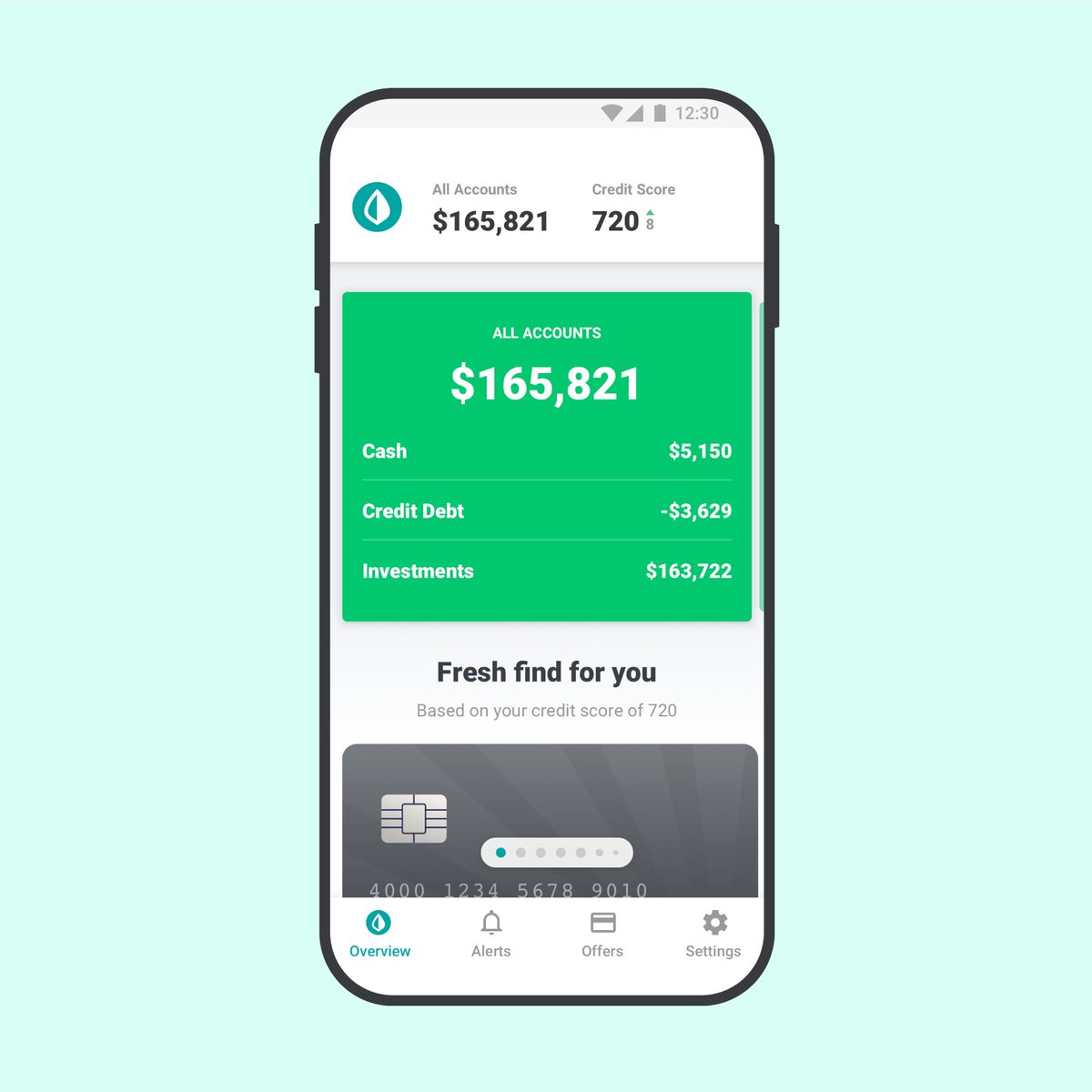
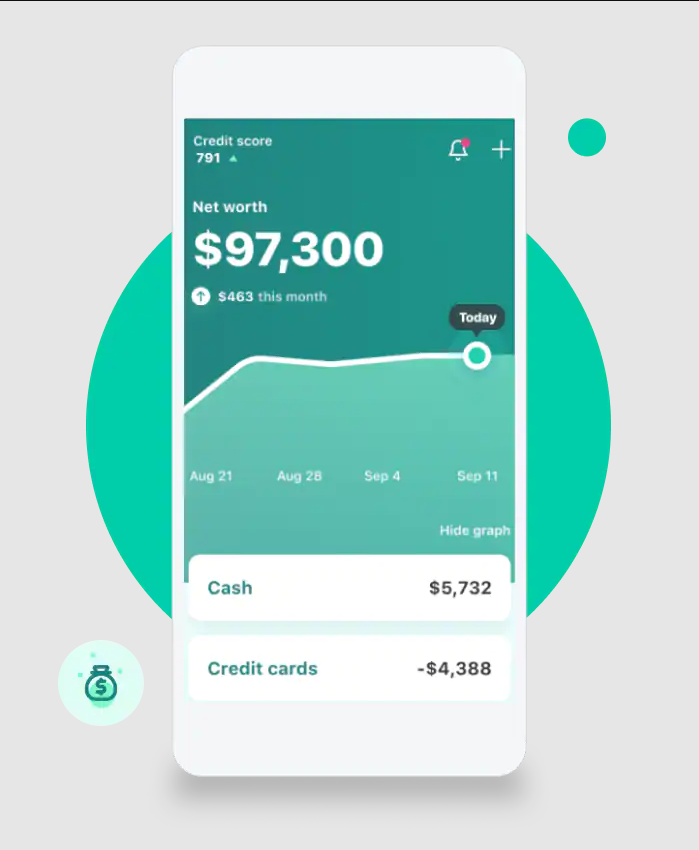
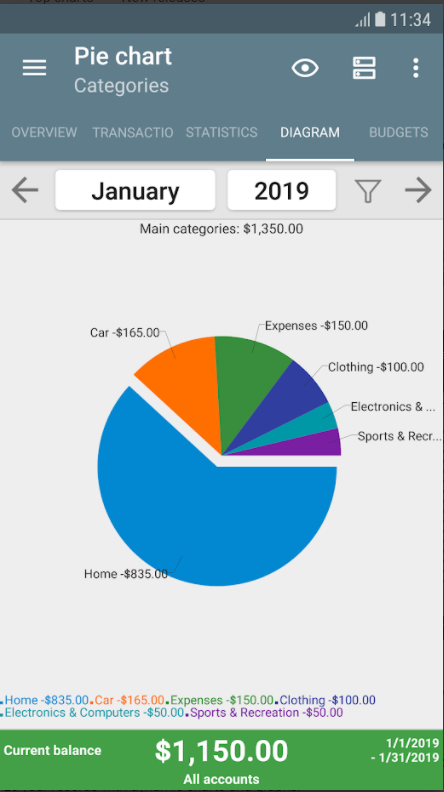


Figure 7:Similar System1: Mint1

Figure 8: Similar System1: Mint2

Above images are of the application mint. After analysis of those application, I found out that they requirement credit card or debit card in order to receive data from the user. In one way this is an advantage since the user does not need to input data as the application and their back account can always stay in sync but this also meant limitation since they could only work with around 10 banks and they were all limited to Canada and USA. Their targeted user base lied to those two specific countries and it will take a lot of time for them to adapt to all the banks in other countries to grow.

Since, this might be better but the targeted user base of the finance application is separate as anyone will have freedom to use it without entering any details and get freedom to later opt in if they want. There will be no partnership with bank because of which the users will have to enter the transaction and amount manually. This will lose integrity than the application mint but it will let my user base to be very scalable very quick without people having to enter their credit card details since not everyone is willing to pull their money at risk for a simple statistical analysis. In addition, specifically to sustaining integrity since not all transactions can be entered in the application there is a separate feature that lets you add correction to the visible amount.



*Figure 9: Similar System2: My Budget Book1*

Another popular system available in Nepal is my budget book. This application is a paid application which provides a platform to manage transactions for the user but it lacks a good UI. This application is mostly targeted for recording the user’s income and expenses along with forecasting the balances. It provides features to statistically view the transactions in order to effectively save up on balance.

Along with a good UI the application also lacks auto categorization feature that will be present in the visioned application. The application will have a naïve bayes algorithm in order to categorize most natural text transactions into well-defined categories. If the algorithm fails to recognize the category then the user can also override the prediction which will add to the training dataset making the algorithm better according the user’s personal preference.

# Bibliography

Agarwal, A., 2019. *support vector machine- formulation and derivation.* [Online]   
Available at: https://towardsdatascience.com/support-vector-machine-formulation-and-derivation-b146ce89f28  
[Accessed 12 12 2021].

Anon., n.d. s.l.:s.n.

Cohen, J., West, S. G. & Aiken, L. S., 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* s.l.:s.n.

GeeksforGeeks, 2021. *Naive Bayes Classifiers.* [Online]   
Available at: https://www.geeksforgeeks.org/naive-bayes-classifiers/  
[Accessed 11 12 2021].

Hastie, T., Tibshirani, R. & Friedman, J., 2008. *The Elements of Statistical Learning.* 2 ed. s.l.:Springer.

Irons, R., 2019. *The Fundamental Principles of Finance.* s.l.:Routledge.

Izenman, A. Z., 2013. *Multivariate Regression. In: Modern Multivariate Statistical Techniques.* New York: Springer.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2021. *Linear Regression: An Introduction to statistical learning.* New York: Springer.

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2021. *Linear Regression: An Introduction to statistical learning.* New York: Springer.

Karasu, S., Altan, A., Sarac, Z. & Hacioglu, R., 2018. *Prediction of Bitcoin prices with machine learning methods using time series data.* s.l., s.n.

Lane, D. M., n.d. *Introduction To Statistics.* Online ed. s.l.:s.n.

Luo, X., 2021. Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal,* pp. 3401-3409.

McCallum, A., 2019. *Graphical Models,* s.l.: s.n.

Nakamoto, S., 2008. Bitcoin: a peer-to-peer electronic cash system.

Nipun Arora, 2019. *Multivariate Linear Regression, MSE, Gradient Descent and Normal Equation.* [Online]   
Available at: https://www.blog.nipunarora.net/ml\_multi\_variate\_linear\_regression/  
[Accessed 10 11 2021].

Piryonesi, S. M. & El-Diraby, E. T., 2020. Role of Data Analytics in Infrastructure Asset Management. *ASCE,* 146(2).

Poongodi, M. et al., 2020. Predicition of the price of Ethereum blockchain cryptocurrency in an industrial finance system. *Computers and Electrical Engineering,* Volume 81.

Rencher, C. A. & Christensen, F. W., 2012. *Multivariate Regression.* 2 ed. s.l.:John Wiley & Sons.

Seltman, H. J., 2018. *Experimental Design And Analysis.* s.l.:s.n.

Shalev-Shwartz, S., Nathan, S. & Cotter, A., 2016. Pegasos: primal estimated sub-gradient solver for SVM. *Mathematical Programming,* 127(1), pp. 3-30.

shalizi, c., 2004. *Quantitative Finance.* s.l.:s.n.

Tschorsch, F. & Scheuermann, B., 2016. Bitcoin and beyond: a technical survey on decentralized digital currencies. *IEEE Commun Surv Tutorials,* Volume 18, pp. 2084-2123.

Vapnik, V. N., 1995. Support-vector networks. *CiteSeerX,* 20(3), pp. 273-297.