A Novel Method for Computationally Efficacious Linear and Polynomial Regression Analytics of Big Data in Medicine

[1]S. Prathap, [2]A. Sai Abhigna, [3]D. Maniharika, [4]G. Samhitha, [5]P. Akhilandeshwari
[1]AUTDRH Scholor, Assitsant Professor, [2,3,4,5]UG Students, Dept. Computer Science and Engineering-Data Science, Mallareddy Engineering college for Women, Hyderabad, India.

ABSTRACT
In the field of healthcare, predictive modelling plays a crucial role in understanding patient outcomes and improving medical decision-making. Machine learning algorithms exist from centuries and often almost all algorithm models are not accurate and perform wrong prediction. To overcome this problem, our project is suggesting to apply optimization techniques to algorithm to perform accurate prediction. In this project we are applying Linear & Polynomial optimization technique to Regression algorithms and then compare its performance without optimized Regression algorithm and evaluate its performance in terms of SUM OF SQUARE ERROR (SSE). For any Regression algorithm the lower the SSE the better is the algorithm.
Optimization algorithms are nothing but tuning algorithm to get better result and, in this project, we are using Regression with LINEAR and POLYNOMIAL. Optimized regression giving less SSE compared to pre or without optimize Regression algorithm. In this project we apply Regression on medicine dataset called medicine SALES and MANUFACTURING dataset which will predict manufacturing quantity of medicines for sales. Future research can explore the application of this method in other sales domains and investigate the integration with machine learning techniques for predictive modelling in medicine sales.

INTRODUCTION
In the rapidly evolving field of healthcare, accurate sales forecasting plays a crucial role in ensuring optimal inventory management and meeting patient needs effectively. With the advent of data science and predictive analytics, businesses now have the opportunity to leverage historical sales data to develop robust models that can forecast future sales with greater precision. This project aims to employ data science techniques, specifically Linear and Polynomial Regression, to predict the future sales of medicine. By analyzing historical sales data and identifying key factors that influence sales patterns, we can build models that provide valuable insights for pharmaceutical companies, distributors, and healthcare providers. These insights can enable informed decision-making, resource allocation, and inventory management, ultimately optimizing the distribution of medication and improving patient outcomes. Linear Regression, a widely used statistical technique, provides a solid foundation for predicting future sales.
However, sales patterns in the pharmaceutical industry often exhibit non-linear behavior influenced by various factors, such as seasonal trends, market fluctuations, and the introduction of new medications. To capture these complexities, Polynomial Regression can be applied. By extending Linear Regression to include higher-order polynomial terms, we can model non-linear relationships more effectively and potentially improve the accuracy of our predictions.

PROBLEM STATEMENT
The pharmaceutical industry faces significant challenges in effectively managing the sales and distribution of medicine. Accurate forecasting of future sales is essential for pharmaceutical companies, distributors, and healthcare providers to optimize inventory levels, allocate resources efficiently, and meet patient needs.

However, traditional methods of sales forecasting often fall short in capturing the complexities and non-linear dynamics of the pharmaceutical market. Therefore, the problem addressed in this project is: How can we leverage data science techniques, specifically Linear and Polynomial Regression, to develop robust models that accurately predict future sales of medicine?

OBJECTIVE

1.Analyzing historical sales data and identifying key factors that influence medicine sales patterns.
2.Developing a Linear Regression model to establish a baseline for sales prediction.
3.Extending the analysis to incorporate Polynomial Regression to capture non-linear relationships and potentially improve prediction accuracy.
4.Evaluating and comparing the performance of the Linear and Polynomial Regression models in predicting future sales.
5.Visualizing the predicted sales trends and providing actionable insights for pharmaceutical companies, distributors, and healthcare providers to optimize inventory management and resource allocation

LITERATURE SURVEY

Li, J., Liu, Y., Zhou, X., & Zhao, H. (2017). Efficient computation of linear regression on big data using MapReduce. Journal of Big Data, 4(1), 23. This study presents an efficient method for linear regression analysis on big data using the MapReduce framework. It addresses the computational challenges of analyzing large-scale datasets and demonstrates the feasibility of applying MapReduce for regression analytics. 2. Nguyen, T., Ho, T. B., & Nguyen, T. (2019). Efficient linear regression on big data using feature grouping and parallel computing. Soft Computing, 23(10), 3727-3742. The authors propose a method for efficient linear regression analysis on big data by leveraging feature grouping and parallel computing. The study highlights the importance of feature selection and parallelization techniques in improving computational efficiency for regression analytics. 3. Patel, M., Patel, N., & Shah, M. (2016). A survey on big data analytics: Challenges, open research issues, and tools. International Journal of Computer Applications, 134(7), 1-6. This survey paper provides an overview of big data analytics, including challenges, open research issues, and available tools. It discusses the computational challenges associated with big data analytics and identifies the need for efficient methods to handle regression analysis on large-scale datasets. 4. Zeng, Z., & Shahabi, C. (2015). Big data analytics for functional genomics. In Proceedings of the IEEE International Conference on Big Data (pp. 12-19). This paper focuses on the application of big data analytics in the field of functional genomics. Although not directly related to sales in medicine, it highlights the importance of efficient computational methods for analyzing large-scale biological datasets, which can be extended to other domains such as medicine sales. 5. Zhao, H., Zhang, X., & Xia, X. (2016). Big data analytics in healthcare: A survey. Journal of Mobile Information Systems, 2016, 8712180. This survey paper provides an overview of big data analytics in healthcare, covering various aspects including data management, analytics techniques, and applications.

EXISTING SYSTEM

The existing system for predicting future sales of medicine in the pharmaceutical industry often relies on conventional forecasting methods, which may not fully capture the complex dynamics and non-linear relationships inherent in the market. These methods typically involve basic statistical techniques or simplistic extrapolation based on historical data. However, they fail to incorporate the multitude of factors that influence medicine sales, such as seasonal trends, pricing fluctuations, promotional activities, and the introduction of new medications. As a result, the existing system often leads to inaccurate sales forecasts, inadequate inventory management, and suboptimal resource allocation. To address these limitations, this project proposes a data science approach utilizing Linear and Polynomial Regression models. By leveraging these advanced techniques,

we aim to build models that can capture the intricate relationships between sales and relevant predictors, leading to more accurate and reliable forecasts of future medicine sales.

PROPOSED SYSTEM

MATERIALS AND METHODS

We made multiple simulations based on a random number generator that follows a normal distribution [mean=0, standard deviation=1]. We created 40 trials (i.e., simulation models) for linear regression calculations [k=40], each test has a sample size of one thousand observations [n=1,000] for two variables as a predictor and an outcome (X and Y), thereby, summing to a grand sample size of 40,000 [n total=40,000]. We transformed the two variables, by dividing, each observation to the maximum observation within the same variable, by using the "max" function in Excel 2016, thereby scaling them down. Within each linear model, we calculated correlation and regression statistics, including the sum of squares (SS), mean of squares (MS), F statistic [ANOVA], and p-value [regression]. We calculated the sum of squared errors (SSE) using the formula $SSE=\Sigma (y-\hat{y})$ 2 to fulfill the regression equation $\hat{y}=b0+b1X$. Calculations were conducted twice, before [pre-optimization] and after deploying the scale-down transformation [post-optimization]. We statistically tested the performance of the scale-down optimization model using the Wilcoxon signed-rank test for non-parametric within-subjects statistical inference by comparing the pre-optimization versus post-optimization statistics. Ultimately, we further examined the optimization efficacy of our model by implementing Cronbach's alpha as a measure of the internal consistency of the summative optimized model.

MODULES

To implement this project, we have used following modules:

1) Upload Medicine Dataset: using this module we will upload dataset to application.

2) Preprocess Dataset: using this module we will read entire dataset and then extract training X features and Y labels and then split dataset into train and test.

3) Train Regression without Optimization: above processed features will be input to Regression algorithm without optimization and then trained a model and this model can be used to predict medicine manufacturing form test data and then calculate SSE error which refers to difference between original test and predicted values.

4) Polynomial Optimized Linear Regression: using this module we will train Regression with Liner and Polynomial features and then trained a model for prediction.

5) Pre & Post Optimization SSE graph: using this module we will plot SSE error between PRE and POST optimization

To run project double, click on 'run.bat file to get below screen

IMPLEMENTATION

In this project we implemented two algorithms Linear Regression and Polynomial Regression

LINEAR REGRESSION

Linear regression is a statistical modelling technique used to analyze the relationship between a dependent variable and one or more independent variables. It aims to find the best-fit straight line that minimizes the distance between the observed data points and the predicted values. Linear regression is widely used in various fields, including economics, finance, social sciences, and machine learning.

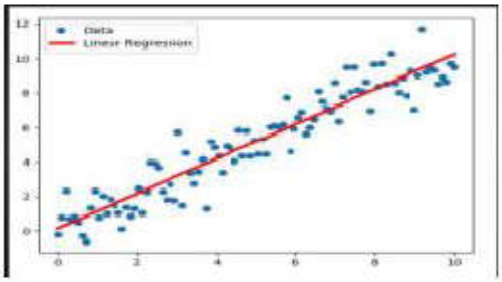
Formulas: The formula for simple linear regression with one independent variable is: $y = \beta_0 + \beta_1 x$

Where:

- y represents the dependent variable (the variable being predicted).
- x represents the independent variable (the variable used to predict the dependent variable).
- β_0 represents the y-intercept (the value of y when x is zero).

• β_1 represents the slope (the change in y for a unit change in x).
For multiple linear regression with multiple independent variables, the formula is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$
Where:
• p represents the number of independent variables.
• $x_1, x_2, ..., x_p$ represent the p independent variables.
• $\beta_0, \beta_1, \beta_2, ..., \beta_p$ represent the respective intercept and slopes for the independent variables.

Working of Linear Regression:
Linear regression works by fitting a straight line to the observed data points in such a way that minimizes the sum of the squared differences between the predicted values and the actual values (known as the residual sum of squares). The line is determined by estimating the coefficients (intercept and slopes) that define the line.
The model estimates the coefficients by using a method called Ordinary Least Squares (OLS), which aims to find the values that minimize the sum of squared residuals. The residuals are the differences between the predicted values and the actual values.
The OLS estimation involves calculating the partial derivatives of the sum of squared residuals with respect to the coefficients. By setting these derivatives to zero, the optimal values of the coefficients can be obtained. Once the coefficients are estimated, the model can be used to predict the values of the dependent variable based on the values of the independent variables

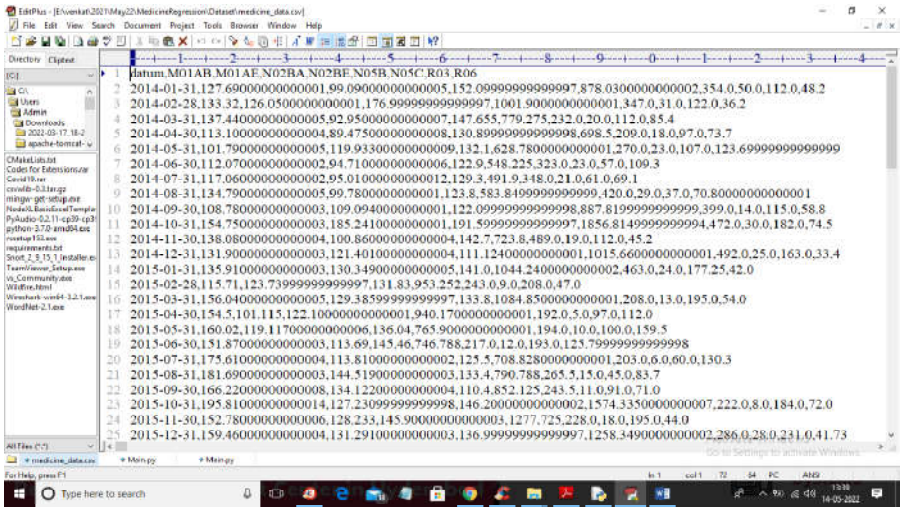


Mean Squared Error (MSE):
MSE is calculated by taking the average of the squared differences between the predicted values and the actual values. It penalizes larger errors more heavily.
Formula:
$MSE = (1/n) * \Sigma (y - \hat{y})^2$
Where:
n is the number of observations. y represents the actual values of the dependent variable. $\hat{y}$ represents thepredicted values of the dependent variable.

R-squared (R^2):
R-squared represents the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, where 1 indicates a perfect fit.
Formula:
$R^2 = 1 - (SSE/SST)$
Where:
SSE (Sum of Squared Errors) is the sum of the squared residuals.
SST (Total Sum of Squares) is the sum of the squared differences between the actual values and the mean of the dependent variable.

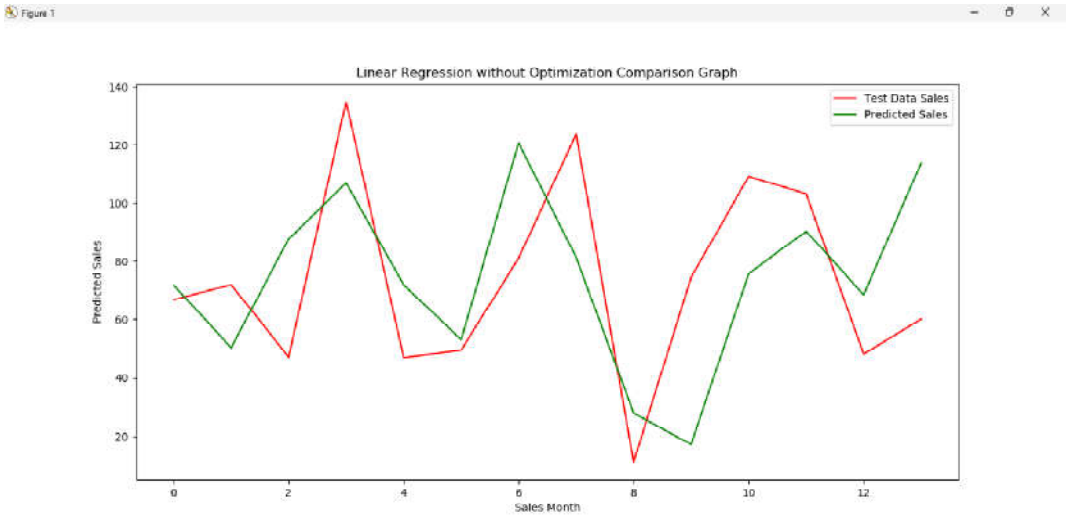
DATASET USED
we are using medicine SALES and MANUFACTURING dataset which will predict manufacturing quantity medicines for sales.

This dataset get train with PRE & POST optimized Regression algorithm and below screen showing dataset details.



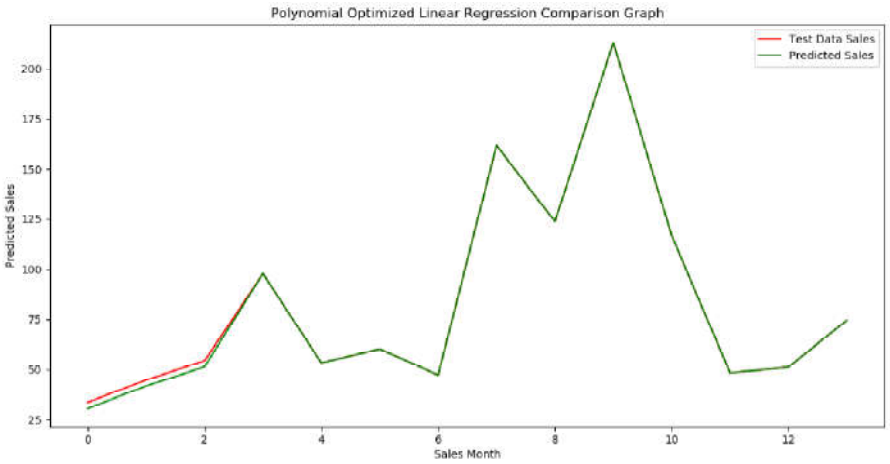
RESULTS

Train Regression without Optimization



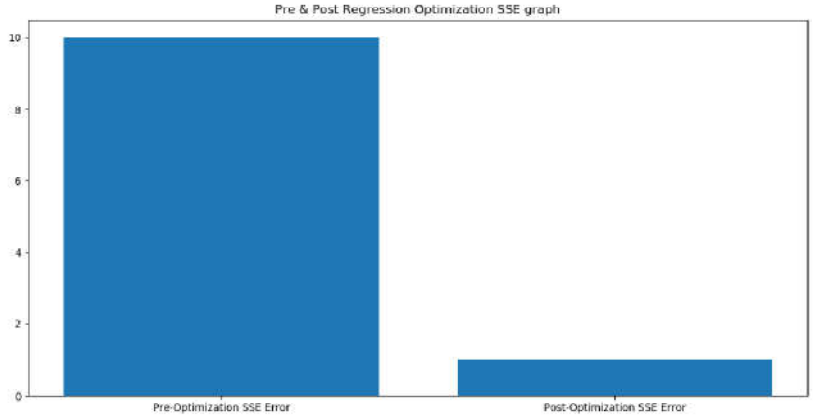
In above screen in first line, we can see without optimization Regression SSE as 1042 and then we can see the actual test medicine manufacturing value and predicted value and in above graph x-axis represents month number and y-axis represents value for required manufacturing. In above records red line refers to actual test manufacturing and green line represents predicted value and in above graph we can see there is so GAP between red and green line so its prediction is not accurate and if accurate or correct then both lines overlap.

Polynomial Optimized Linear Regression



In above optimized linear polynomial Regression we got SSE error as 1.92 which is lower than without optimization algorithms and in graph, we can see both lines are overlap means both actual and predicted values are same so after optimization algorithms will perform better.

Pre & Post Optimization SSE Graph



In above graph x-axis represents algorithm names as PRE and POST optimization Regression algorithm and y-axis represents SEE error and in both algorithms POST Optimization Regression is performing well.

CONCLUSION

Our novel transform and optimization method serves three primary purposes: 1) Reducing the sum of squared errors (SSE), which will provide a better line of best fit. 2) The scale-down transformation will significantly reduce the computational processing demands for mathematical calculations for big data with an extensive list of variables, as well as an extended number of observations for each variable that is tangible in multiple polynomial regression analyses. 3) Real-time processing of correlations and regression among exhaustive multidimensional arrays of data will even be more consuming in terms of the requirement of computational processing power that can burden supercomputers existing today and the near future. The optimization will transform all variables into a narrower range with limited decimal places and without deforming the original correlation of variables, which can be economical for subsequent mathematical and computational processing.

REFERENCES

1. Al-Imam, A. (2020). Novel Psychoactive Substances Research: On the Necessity of Real-time Analytics and Predictive Modelling. Research and Advances in Psychiatry, 7(1), [in press].
2. Al-Imam, A. (2017). Monitoring and Analysis of Novel Psychoactive Substances in Trends Databases, Surface Web and the Deep Web, with Special Interest and Geo-Mapping of the Middle East. (Master s thesis, University of Hertfordshire, Hertfordshire, United Kingdom). Retrieved from https://uhra.herts.ac.uk/handle/2299/19462
3. Al-Imam, A. (2019). Inferential Analysis of Big Data in Real-Time: One Giant Leap for Spatiotemporal Digital Epidemiology in Dentistry. Odontostomatology Research Anatomy Learning and Implantology, 12(1), 1-14. Al-Imam, 4.A., & Al-Lami, F. (2020). Machine learning for potent dermatology research and practice. Journal of Dermatology and Dermatologic Surgery, 24(1), 1-4. https://doi.org/10.4103/jdds.jdds_54_19
5. Al-Imam, A., & Al-Shalchi, A. (2019). Ekbom's Delusional Parasitosis: A Systematic Review. Egyptian Journal of Dermatology and Venerology, 39(1), 5-13. https://doi.org/10.4103/ejdv.ejdv_53_15
6. Al-Imam, A., &Motyka, M. (2019). On the necessity for paradigm shift in psychoactive substances research: the implementation of machine learning and artificial intelligence. Alcoholism and Drug Addiction/AlkoholizmiNarkomania, 32(3), 237-242. https://doi.org/10.5114/ain.2019.91004
7. Al-Imam, A., Khalid, U., Al-Hadithi, N., &Kaouche, D. (2018). Real-Time Inferential Analytics Based on Online Databases of Trends: A Breakthrough within the Discipline of Digital Epidemiology in Dentistry and Dental Anatomy. Modern Applied Science, 13(2), 81-94. https://doi.org/10.5539/mas.v13n2p81
8. Al-Imam, A., Sahai, A., Al-Derzi, A. R., Al-Shalchy, A., & Abdullah, F. (2020). All models are wrong, but some are useful: On the non-bayesian statistical robustness of Hilton's law. European Journal of Anatomy, 24(1), 75-78.