Significance of Matrix Rank and Pseudo-Inverse Techniques: Regression and Classification Approaches for Predictive Modelling in Stock Price Dynamics and Change Percentage Trends

Vansh Kushwaha

Department of Computer Science Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India

BL.EN.U4CSE21216@bl.students.amrita.edu

Ve Ram Akathya

Department of Computer Science Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India

BL.EN.U4CSE21217@bl.students.amrita.edu

Yadukul S

Department of Computer Science Engineering, Amrita Vishwa Vidyapeetham, Bangalore, India

BL.EN.U4CSE21222@bl.students.amrita.edu

***Abstract* – This report presents comprehensive solutions for matrix properties and linear algebra challenges, analyzing data from the "Purchase Data" worksheet and "IRCTC Stock Price" dataset. Through regression and classification techniques, it delves into the essence of rank importance while distinguishing between these methodologies. Utilizing libraries for matrix rank determination, lambda functions, and linear algebraic functions for classification and regression discussions, the report further incorporates central tendencies for comprehensive data description. This approach leads to optimal outcomes in terms of mean, variance, and statistical evaluations.**

Principal phrases:* *Rank, Classification, Regression, Conditional Probability, Data Analytics, Mean, Data visualization, data preprocessing

Introduction

This study seeks to comprehensively investigate multiple facets of the provided dataset, covering dimensions such as vector space dimensionality, vector count, feature matrix rank, and cost calculation through pseudo-inverse methods. The initial phase involves extracting pertinent data and organizing it into two distinct matrices, labeled matrix A and matrix C, following the 'Ax = C' notation. The dimensionality of the vector space is intrinsically linked to the number of features associated with each product, while the vector count aligns with the total products present in the dataset. Through the computation of the rank of matrix A, valuable insights are unveiled concerning the independence of constituent features.

Simultaneously, the analysis of the 'IRCTC stock price' dataset delves into multifaceted aspects encompassing mean and variance, offering a holistic view of central tendencies and data variability. The preliminary stages encompass data preparation and preprocessing for the stock price dataset, enhancing the ability to deduce probabilities tied to potential stock loss scenarios through percentage change data. This in-depth approach allows for a comprehensive understanding of the risk landscape connected with investment decisions.

In parallel, the realm of supervised machine learning assumes prominence, characterized by its utilization of labeled data to train algorithms for classification and precise result predictions. This technique proves particularly valuable for addressing real-world challenges at scale, such as categorizing spam emails into dedicated folders. Notable techniques employed in supervised learning encompass linear regression, logistic regression, classification, and others.

Amidst the contemporary landscape, the extraction of meaningful insights from complex datasets is of paramount importance. Data Analytics, the practice of dissecting raw data to glean informative inferences, is further enhanced through the integration of machine learning algorithms. This synergy is known as machine learning for analytics and aids in the evaluation of data and the revelation of insights that drive informed decision-making and improve business outcomes.

Leveraging a toolkit of linear algebra methods, mathematical tools are adeptly wielded to tackle challenges involving vectors and matrices. These methods prove versatile, finding application in classification tasks where they facilitate feature extraction from input data, contributing to model training capable of predicting labels for new data points. Mean and variance play a pivotal role in understanding central dispersion, with the mean representing the average of a set of numbers and variance signifying the average squared deviations from the mean.

Incorporating visual representations into the analytical process, plotting graphs showcasing change percentage and days becomes a powerful tool. In the context of machine learning, this is achieved through invoking the 'plot()' function and providing the change percentage

and days data as inputs, effectively translating intricate patterns into visual insights.

PROBLEM DESCRIPTION

At the core of this report lies the primary challenge of transforming tabular data from the "Purchase Data" worksheet into matrices A and C, employing the AX= C framework. Subsequently, a spectrum of analytical tasks unfolds, encompassing the computation of Matrix A's rank, estimation of product costs through pseudo-inverse methods, prediction of product costs via a model vector X, and the formulation of a client classification model based on their purchasing patterns.

Simultaneously, the study addresses pivotal questions surrounding the "IRCTC Stock Price" dataset. It undertakes the computation of essential statistics like mean and variance for stock prices, conducts comparative analysis of means for Wednesdays and April, evaluates the probabilities of encountering profit or loss scenarios, and delves into the domain of conditional probability.

DATA DESCRIPTION

The dataset encompasses a range of input attributes, including customer purchases like 'candies', 'mangoes', 'milk packets', and 'payments', with the target attribute being customer purchases. In parallel, the 'IRCTC stock price' dataset sheds light on stock movements across a specific timeframe. Its input attributes comprise 'month', 'day', 'high', 'low', and 'volume', while the target attributes encompass 'price' and 'change%'.

Further into the process, data preprocessing emerges as a pivotal step, encompassing data refinement, transformation, and integration to render it primed for analysis. The overarching objective of data preprocessing remains consistent: elevating data quality and adapting it to the precise requirements of data mining endeavors.

To initiate data preprocessing, the 'data.dropna' command is adeptly employed, facilitating the removal of null and missing values. Additionally, an alternative approach involves handling null values by imputing central tendencies like mean, median, or mode, preserving the dataset's integrity.

Categorically, we delve into two distinct datasets: "Purchase Data" and "IRCTC Stock Data." The former takes the form of tabular data in Excel, with columns like 'Customer', 'Candies', 'Mangoes', 'Milk Packets', and 'Payment'. The dimensions fluctuate with varying rows and columns. In parallel, the "IRCTC Stock Data" also assumes a tabular Excel format, featuring columns such as 'Date', 'Month', 'Day', 'Price', 'Open', 'High', 'Low', 'Volume', and 'Chg%'. Similar to the previous dataset, these dimensions vary in terms of rows and columns.

The overarching analysis objectives include calculating the rank, dimension, and pseudo-inverse; determining mean and variance; estimating the probability of incurring losses based on 'Chg%'; computing overall probability and conditional probability; generating a scatter plot depicting 'Chg%' data against weekdays for intuitive visualization; and executing scatter plotting techniques. This comprehensive approach equips us with profound insights into the data's intricate patterns and behaviors.

METHODOLOGY

Study 1: Classification Model Development

Importance of Rank of Observation Matrix in Classification:

The rank of an observation matrix is a critical factor in model building for classification. In machine learning, linear algebraic operations are often employed, and the rank of the observation matrix impacts model effectiveness, stability, and generalization.

1. Redundancy: High multicollinearity, which occurs when features are highly correlated, can reduce the rank of the matrix. A low rank suggests feature redundancy and the potential for linear combinations among features.

2. Stability: When the matrix rank is low, the model solutions can become sensitive to minor changes in the data, reducing its reliability.

3. Regularization: Regularization techniques are essential when the matrix has full rank but a small sample size. Regularization helps prevent overfitting and improves model generalization.

Data Collection/Preprocessing and Feature Selection:

The dataset was obtained from "LabSession1_Purchase.xlsx," loaded using the pandas library in Google Colab. Missing values were handled by dropping columns with missing data. Features "Candies (#)," "Mangoes (Kg)," and "Milk Packets (#)" were selected due to their potential influence on payment amounts.

Target Variable and Rank Calculation:

The target variable, "Payment (Rs)," was chosen as the dependent variable. The rank of the feature matrix was calculated using 'np.linalg.matrix_rank,' revealing the dimensionality of the feature space and potential feature correlations.

```
[5]  rankA = la.matrix_rank(A)
     print("The rank of matrix A calculated is ",rankA)

     The rank of matrix A calculated is  3
```

Matrix Calculation:

A matrix equation was formulated using the selected features and the target variable. The inverse of the feature matrix, denoted as A-1, was used to calculate the coefficients describing the relationship between selected features and payment amounts.

Categorization of Customers:

Customers were categorized based on payment amounts into "Rich" or "Poor" categories, using a threshold of 200 Rs.

```
[7]  val['rich/poor'] = np.where(val['Payment (Rs)'] > 200,'Rich','Poor')
     df_cleaned = val.dropna(axis=1,how='any')
     print(df_cleaned)
```

Study 2: Analysis of IRCTC Stock Prices

Introduction:

This study analyzes stock market data from "LabSession1_Stock.xlsx" to understand stock price distributions, variations on specific days (Wednesdays), and probabilities of profit and loss.

Data Acquisition:

Stock market data was sourced from "LabSession1_Stock.xlsx" and loaded into Google Colab using the pandas library.

Statistical Calculations:

Mean and variance of stock prices were computed to gain insights into the distribution and volatility of stock prices.

```
mean = statistics.mean(df['Price'])
var = statistics.variance(df['Price'])

print("The mean value of Price column is ", mean , "and the variance value is ", var)
```

```
The mean value of Price column is  1560.663453815261 and the variance value is  58732.365352539
```

Analysis of Specific Days (Wednesday):

Data was segmented based on Wednesdays, and the mean of stock prices for Wednesdays was calculated and compared to the population mean, offering insights into Wednesday-specific variations.

Probability Analysis:

The dataset was analyzed to identify instances of stock price declines (negative percentage changes). The probability of making a loss in the stock market was computed. Additionally, the probability of making a profit specifically on Wednesdays was calculated.

```
print("Probability of Making a Profit given that it is Wednesday: ",profit_count/wednesday_count)
```

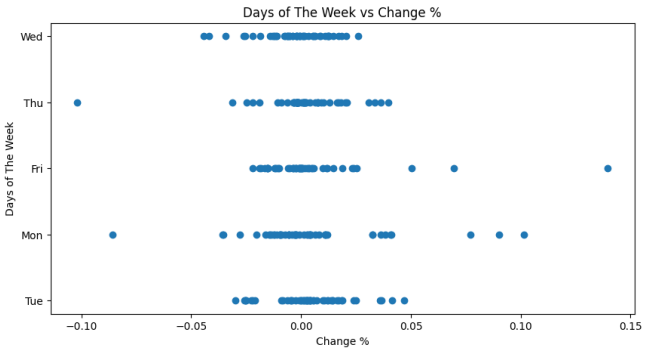
```
Probability of Making a Profit given that it is Wednesday:  0.42
```

```
wednesday_count = df['Chg%'].loc[df['Day'] == 'Wed'].count()
profit_count = df['Chg%'].loc[(df['Day'] == 'Wed') & (df['Chg%'] >0)]
loss_count = df['Chg%'].loc[(df['Day'] == 'Wed') & (df['Chg%'] <0)]
loss_count = loss_count.count()
profit_count = profit_count.count()

print("Probability of a Day Being a Wednesday: ",wednesday_count/count_total)

print("Probability of Making a Profit on Wednesday: ",profit_count/count_total)

print("Probability of Making a Loss on Wednesday: ",loss_count/count_total)
```

```
Probability of a Day Being a Wednesday:  0.20080321285140562
Probability of Making a Profit on Wednesday:  0.08433734939759036
Probability of Making a Loss on Wednesday:  0.11646586345381527
```

Data Visualization:

A scatter plot was created to visualize the relationship between the day of the week and daily percentage changes in stock prices. This visualization provides insights into stock price volatility across different days.



RESULT & ANALYSIS

This report presents the results of the analysis performed on two datasets: the purchase dataset and the IRCTC stock price dataset. The analysis involved slicing datasets, applying matrix algebra principles, calculating probabilities, and creating visualizations to gain insights into the data.

The purchase dataset was subjected to matrix algebra techniques, including rank calculation and pseudo-inverse computation. These techniques allowed for precise data manipulation.

An analysis of the dataset revealed that the probability of a day falling on Wednesday is higher, indicating that Wednesday is a common day for purchases.

The analysis of the stock price data indicated that the sample means for Wednesdays and the month of April closely align with the population mean. This suggests that the stock's performance remains relatively stable during these periods.

The calculation of the probability of making a loss highlights the inherent risk associated with trading IRCTC stock. This information is crucial for risk assessment.

The computed probabilities of making a profit on Wednesdays and the conditional probability of making a profit on Wednesdays provide valuable insights into the potential profitability of trading on Wednesdays.

Investors can use this information to make informed decisions based on the day of the week.

A scatter plot depicting the daily percentage changes (Chg%) in stock prices against the day of the week was created. This visualization assists in identifying trends, outliers, and potential patterns that might influence trading strategies.

Conclusion

In the comprehensive analysis of both the "Purchase Data" and "IRCTC Stock Price" databases, we have unearthed valuable insights and knowledge. We systematically explored the data structures, cost estimations, and customer categorization within the purchase dataset. Simultaneously, our investigation into the stock price dataset revealed trends, associated risks, and profit prospects for informed stock research.

These insights serve as critical assets, empowering organizations to fine-tune their strategies and enabling investors to make well-informed and calculated judgments. The application of this knowledge holds the potential to pave the way for success in their respective fields.