Mining Massive Data Sets Midterm Report

1st 522H0036 - Luong Canh Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0036@student.tdtu.edu.com

2nd 522H0092 - Cao Nguyen Thai Thuan
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0092@student.tdtu.edu.com

3rd 522H0075 - Tang Minh Thien An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0075@student.tdtu.edu.com

4th 522H0167 - Truong Tri Phong
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
522H0167@student.tdtu.edu.com

5th Instructor: Nguyen Thanh An
Faculty of Information Technology
Ton Duc Thang University
Ho Chi Minh City, Vietnam
nguyenthanhan@tdtu.edu.com

***Abstract*—This project implements and evaluates key techniques in mining massive datasets. It covers hierarchical agglomerative clustering of string shingles using Jaccard distance; PySpark-based linear regression for gold price prediction; CUR decomposition for feature dimensionality reduction on gold price data; and PageRank analysis of the `it.tdtu.edu.vn` web graph using PySpark. The work demonstrates practical applications and provides insights into processing large-scale data.**

I. Introduction

The increasing volume of data requires efficient mining techniques. This project implements and analyzes four core algorithms: (1) hierarchical agglomerative clustering for non-Euclidean text data, using 4-shingles and Jaccard distance on alphabetical strings; (2) PySpark-based linear regression to predict Vietnamese gold prices from historical data; (3) CUR decomposition to reduce the dimensionality of gold price features (from 10 to 5) and assess its impact on regression; and (4) PageRank, implemented in PySpark, to identify influential pages within the `it.tdtu.edu.vn` web graph. Python and PySpark are utilized throughout. This report details the methodologies, implementations, and experimental results for each task.

II. First Task: Hierarchical clustering in non-Euclidean spaces

III. Second Task: Linear Regression – Gold price prediction

This task focused on predicting Vietnamese gold prices using a linear regression model implemented in PySpark. The objective was to transform historical time-series data into a suitable format for regression, train a model, and evaluate its predictive performance.

A. Overview of Linear Regression

Linear Regression is a supervised learning algorithm that models the linear relationship between a continuous target variable (y) and one or more independent predictor variables (features $\mathbf{x}$). The goal is to find an optimal linear function that best predicts y given $\mathbf{x}$.

For a single feature x, the model is:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

With multiple features $x_1, x_2, \ldots, x_p$, it extends to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \tag{2}$$

where β_0 is the intercept, β_j are the feature coefficients (weights), and ϵ is the error term. The coefficients are typically learned by minimizing a loss function, such as Mean SquaredError (MSE), often using optimization algorithms like L-BFGS. Key assumptions include linearity, independence of errors, and homoscedasticity. This project applies linear regression to predict gold prices based on historical price features.

B. Data Preparation

- **Dataset:** The primary data source was `gold_prices.csv` (2009/08/01 to 2025/01/01), read into a PySpark DataFrame.
- **Feature Engineering:** For each target date t, features were the respective 'Buy Price' or 'Sell Price' values from the 10 consecutive preceding days. PySpark's `Window` functions and `lag` operation were used, followed by `VectorAssembler` to create feature vectors (e.g., `Previous Buy Price(s)`). 4000 samples were generated (`random_state=38`).
- **Data Splitting:** The generated DataFrame was randomly split into training (70%) and testing (30%) sets (`seed=2`).

C. Model Implementation and Training

Two separate linear regression models (`pyspark.ml.regression.LinearRegression`) were developed: one for 'Buy Price' and one for 'Sell Price', using their respective 10-day historical price vectors as features and the current price as the label. Models were

configured with the `'l-bfgs'` solver and trained on the 70% training subset.

D. Experimental Results and Evaluation

The performance of the trained models was evaluated on both training and testing sets. Key metrics are summarized in Table I. The R^2 values (consistently > 0.999), low RMSE/MAE, and high Explained Variance scores indicate strong predictive accuracy and good generalization to unseen data, with no significant overfitting observed.

TABLE I
LR PERFORMANCE METRICS FOR GOLD PRICE PREDICTION.

Model	Data Set	RMSE	MSE	R^2	MAE	Expl. Var.
Buy Price	Training	0.3232	0.1045	0.9995	0.1450	225.7249
	Test	0.2975	0.0885	0.9996	0.1356	226.5322
Sell Price	Training	0.3062	0.0938	0.9996	0.1406	240.1171
	Test	0.2913	0.0848	0.9996	0.1319	240.6501

E. Visualizations

1) Loss History During Training:

Line chart (Fig. 1) illustrated the objective function value per iteration, showing rapid convergence for 'Buy Price' model.

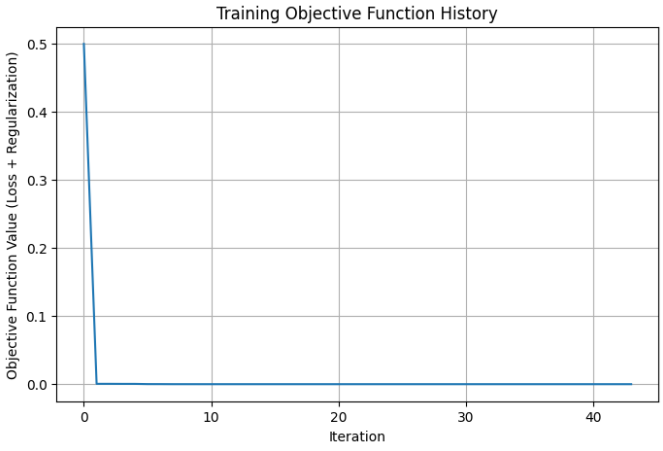

Fig. 1. Loss History of Buy Price Prediction Model

2) Performance Comparison:

Bar charts (Fig. 2 nd Fig. 3) contrasted evaluation metrics (RMSE, MSE, R^2, MAE, Explained Variance) between training and testing sets, visually confirming robust generalization.

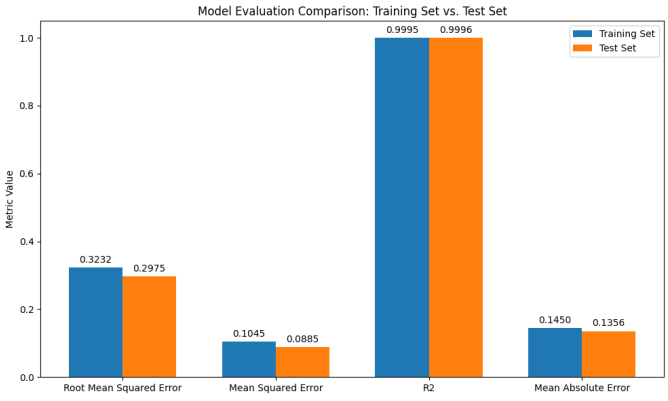

Fig. 2. 'Buy Price' Model Performance (Training vs. Test Set).

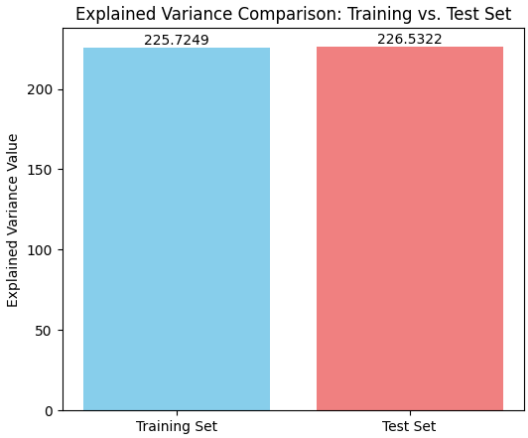

Fig. 3. 'Buy Price' Model Explained Variance Performance (Training vs. Test Set).

IV. Third Task: CUR – Dimensionality Reduction

V. Fourth Task: PageRanking – the Google algorithm

VI. Contribution

The following table represents the contribution of each member, note that whichever member handles whichever task will also write the report for that task.

TABLE II
MEMBER CONTRIBUTIONS

ID	Member	Contribution	Progress
522H0036	Luong Canh Phong	Task 2 and Handling Report	100%
522H0092	Cao Nguyen Thai Thuan	Task 4 and Report Support	100%
522H0075	Tang Minh Thien An	Task 3	100%
522H0167	Truong Tri Phong	Task 1	100%

VII. Self-evaluation

The following table is our self-evaluation on our tasks:

VIII. Conclusion

References

TABLE III
SELF-EVALUATION

Task	**Task Requirements**	**Completion Ratio**
Task 1	Hierarchical clustering in non-Euclidean spaces	100%
Task 2	Linear Regression – Gold price prediction	95%
Task 3	CUR – Dimensionality Reduction	90%
Task 4	PageRanking – the Google algorithm	100%
Task 5	Report	100%