A Report on PySpark Assignment

Submitted By:

Nirmala Regmi

Submitted To:

Ishant Gupta

1. Introduction

This report provides an analysis and machine learning workflow using PySpark on a company dataset. The main goals were to clean and explore the data, visualize insights, and build a predictive model for estimating current_assets.

2. Dataset Description

The dataset comprises various attributes about companies, including:

- **General Information**: company_number, company_type, jurisdiction
- **Financial and Employment Details**: current_assets, average_number_employees_during_period
- **Status Indicators**: company_status, next_accounts_overdue, confirmation_statement_overdue

Sample Structure

The dataset includes columns such as:

- company_number (string)
- company_type (string)
- current_assets (double)
- incorporation_date (string, converted to date)
- company_status (string)

3. Data Loading and Initial Exploration

Objective

Load the data into PySpark, confirm the structure, and conduct a preliminary exploration to understand the dataset.

Summary

- The dataset was loaded successfully, with 5,428,900 rows.
- Preliminary exploration displayed the schema and sample data.

Importance

Initial data exploration helps ensure the dataset is correctly structured and ready for cleaning and analysis.

4. Data Cleaning and Transformation

Objective

Prepare the dataset for analysis by addressing missing values, removing duplicates, and converting data types.

Steps and Key Results

- **Duplicate Removal**: Duplicates were dropped to maintain unique entries.
- **Handling Missing Values**: Missing numeric columns were filled with 0, while text columns were filled with "Unknown".
- **Data Type Conversion**: Certain columns were converted to appropriate data types for analysis.

Importance

Data cleaning ensures the dataset's completeness and accuracy, forming the basis for reliable analysis and modeling.

5. Data Analysis and Visualizations

Objective

Explore the dataset using Spark SQL and generate visualizations to highlight key insights.

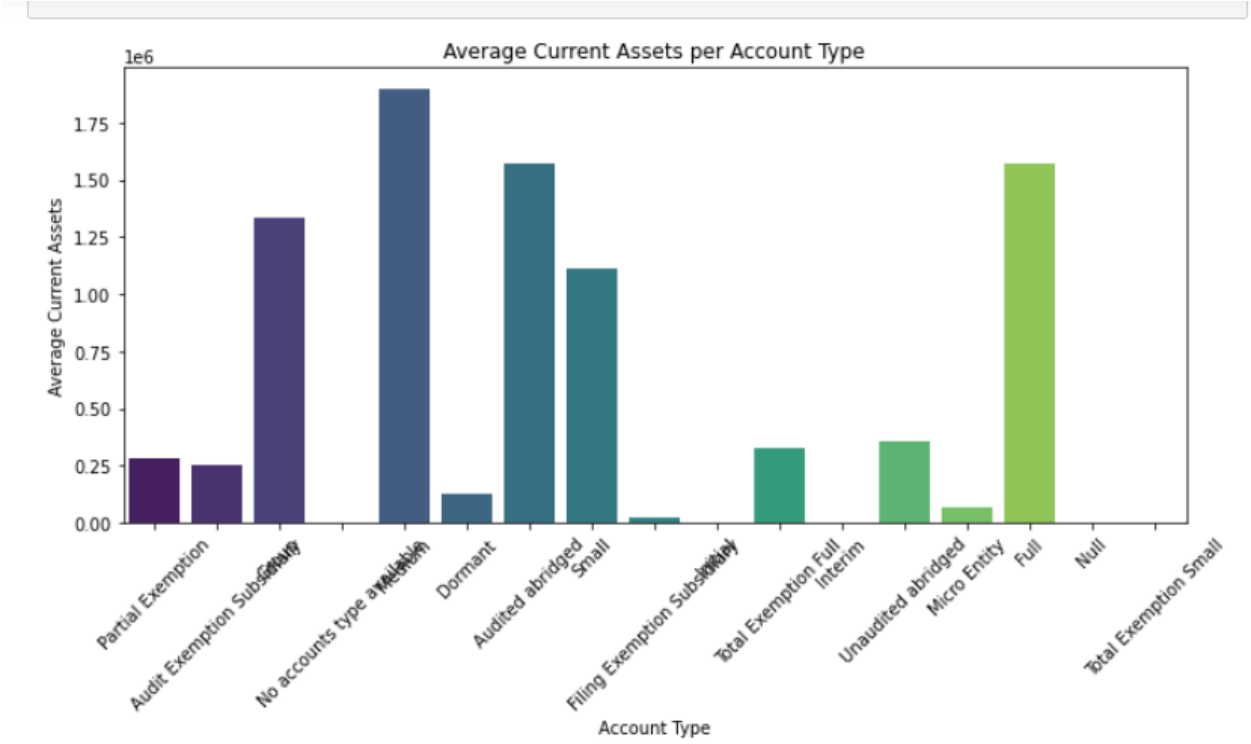
Analysis Results

1. **Aggregation**:
    - The average current_assets was approximately 184,785.89, with a high standard deviation indicating variability among companies.
2. **Grouping by company_status**:
    - Active companies had an average current_assets of around 203,061.28, while dissolved companies had a much lower average of 24,222.90.
3. **Time-Based Analysis**:
    - The number of company incorporations showed consistent growth over time, particularly peaking in recent decades.

Visual Insights

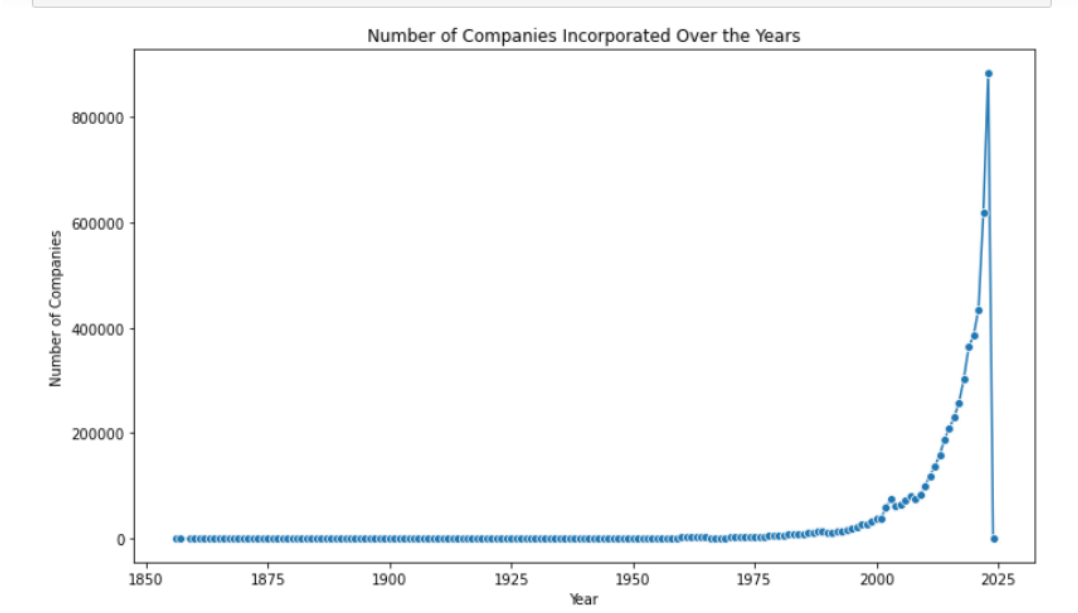
1. **Average Current Assets by Account Type**:

- A bar chart highlighted differences in average current_assets across various account types, showing significant variations between account categories.



2. **Company Incorporation Trends**:

- A line chart illustrated the trend of company incorporations over the years, confirming a steady increase.



Importance

Visualizations make it easier to understand data patterns and convey findings clearly. They are vital for communicating insights effectively to a wider audience.

6. Machine Learning Model

Objective

Build a regression model to predict current_assets using selected company attributes.

Summary of Results

- **Feature Selection**: Key features included owners, officers, and average_number_employees_during_period.
- **Initial Model Performance**: The root mean squared error (RMSE) was approximately 11,997,463.59, indicating the initial model's predictive capabilities.

Importance

A regression model enables financial forecasting and risk assessment by predicting asset levels based on company attributes.

7. Model Tuning and Evaluation

Objective

Enhance the model's performance through cross-validation and hyperparameter tuning.

Key Findings

- Cross-validation was conducted using a parameter grid and multiple folds to optimize the model.
- The **final RMSE** remained around 11,997,463.59, suggesting further refinements or additional features may be needed to improve accuracy.

Importance

Model tuning improves generalizability and predictive performance by optimizing hyperparameters and evaluating the model across different data splits.

8. Conclusion

The analysis provided key insights into the dataset, such as:

- **Variability in current_assets** based on company status.
- **Steady growth in company incorporations** over time.
- **Predictive modeling** showcased the potential for estimating current_assets, although further model refinements are suggested.