CSP 571 Project proposal and project outline

Company Financials for Predictive Insights and Economic Trends

Team Members:

CWID	Names
A20543155	Gangili Sai Charan
A20545853	Neeraj Vardhan Buneeti
A20539719	Poojith Reddy Annachedu

Project Proposal:

Description of the Project:

This project achieves a systematic analysis of the "Company Financials Dataset". Finding useful information about the financial standing and performance trends of listed companies is the main objective of the research. We hope to gain insight into the basic causes of business success as well as any potential red flags of financial trouble through this analysis.

Research Goal:

- To pinpoint the most important financial indicators that indicate a business's potential for expansion and success.
- To create a predictive model using historical data that can project a company's financial outlook.

Research Questions:

1. What financial indicators are most reliable for predicting future profitability and growth of a company?

2. Is it possible to estimate the financial condition of a business for the coming financial year using historical data?

3. In terms of growth patterns and financial stability, how do various industries compare?

4. Are there any patterns or irregularities that could indicate changes in the economy or disturbances in the market?

Proposed Methodology/Approach:

The analysis will be conducted in several stages:

1. **Data Pre-processing**: Preparing the dataset for analysis by cleaning, normalizing, and structuring it.

2. **Exploratory Data Analysis (EDA)**: Summarizing the shape, dispersion, and central tendencies of the dataset's distribution using statistical tools.

3. **Feature Selection**: Choosing important financial indicators by using methods such as feature importance analysis and correlation matrices.

4. **Model Development**: The process of developing predictive models involves utilizing different statistical and machine learning algorithms, such as regression analysis, decision trees, random forests, and neural networks.

5. **Model Evaluation and Tuning**: Using cross-validation, evaluate the model's performance and modify its parameters to increase precision.

6. **Industry Comparison**: Comparing and contrasting industries to assess the financial health of various sectors.

7. **Trend Analysis**: Examining time-series data for any cyclical or long-term financial trends.

Metrics for Measuring Analysis Results: -

- R^2 score (Coefficient of Determination): The percentage of the dependent variable's variance that can be predicted based on the independent variables.
- Mean Absolute Error (MAE): Used to gauge how closely the results match the forecasts.
- Root Mean Squared Error (RMSE): This statistic calculates the square root of the errors' average squared.
- Accuracy Score represents the proportion of all correctly predicted outcomes in classification problems (e.g., estimating a company's profitability).
- F1 Score: Harmonic mean of recall and precision; especially helpful in cases where classes are unbalanced.
- AUC-ROC Curve, is a metric used to evaluate binary classification issues.

2 Project Outline

Literature Review and Related Work

Existing Projects and Studies: Research and documentation of existing projects or studies that have utilized similar datasets for financial analysis. This would include reviewing the methodologies and findings of these projects to understand the current state of financial data analysis.

Academic References: Compilation of academic papers and scholarly articles that provide a foundation for financial data analysis, predictive financial modelling, and sectoral financial comparison.

Relevant Articles and Resources: Collection of articles, blog posts, and other resources that offer insights into financial data analysis techniques and industry-specific financial trends.

Data Sources and Reference Data

Dataset Details:

- Kaggle - Company Financials Dataset https://www.kaggle.com/datasets/atharvaarya25/financials/data
- Us Security financial statement data set https://www.sec.gov/files/aqfs.pdf

Column	Datatype	Description
Segment	Object	Categorization of the market segment to which the product sales data applies.
Country	Object	The country where the product was sold.
Product	Object	The specific product that the sales data is recorded for.
Discount band	Object	The range or category of discount that was applied to the sales.
Units sold	Int64	The quantity of product units sold.
Manufacturing price	Object	The cost to manufacture one unit of the product.

Sale price	Object	The price at which one unit of the product is offered for sale to the customer.
Gross sales	Object	The total sales revenue before any discounts are applied.
Discounts	Object	The total amount of discounts given on the gross sales.
Sales	Object	The final sales revenue after applying discounts.
COGS	Object	Cost of Goods Sold - the direct costs attributable to the production of the products sold.
Profit	Object	The financial gain calculated as the difference between the sales and the cost of goods sold.
Date	Date	The specific date when the sale was recorded.
Month Number	Int64	The numerical representation of the month when the sale occurred (1-12).
Month Name	Object	The name of the month when the sale was recorded.
Year	Int64	The year when the sale occurred.

Company Financial Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
674	Government	Mexico	Paseo	High	$2,535.00	$10.00	$7.00	$17,745.00	$2,661.75	$15,083.25	$12,675.00	$2,408.25	01-04-2014	4	April	2014
675	Government	Mexico	Paseo	High	$2,851.00	$10.00	$350.00	$9,97,850.00	$1,49,677.50	$8,48,172.50	$7,41,260.00	$1,06,912.50	01-05-2014	5	May	2014
676	Midmarket	Canada	Paseo	High	$2,559.00	$10.00	$15.00	$38,385.00	$5,757.75	$32,627.25	$25,590.00	$7,037.25	01-08-2014	8	August	2014
677	Government	United States of America	Paseo	High	$267.00	$10.00	$20.00	$5,340.00	$801.00	$4,539.00	$2,670.00	$1,869.00	01-10-2013	10	October	2013
678	Enterprise	Germany	Paseo	High	$1,085.00	$10.00	$125.00	$1,35,625.00	$20,343.75	$1,15,281.25	$1,30,200.00	$(14,918.75)	01-10-2014	10	October	2014
679	Midmarket	Germany	Paseo	High	$1,175.00	$10.00	$15.00	$17,625.00	$2,643.75	$14,981.25	$11,750.00	$3,231.25	01-10-2014	10	October	2014
680	Government	United States of America	Paseo	High	$2,007.00	$10.00	$350.00	$7,02,450.00	$1,05,367.50	$5,97,082.50	$5,21,820.00	$75,262.50	01-11-2013	11	November	2013
681	Government	Mexico	Paseo	High	$2,151.00	$10.00	$350.00	$7,52,850.00	$1,12,927.50	$6,39,922.50	$5,59,260.00	$80,662.50	01-11-2013	11	November	2013
682	Channel Partners	United States of America	Paseo	High	$914.00	$10.00	$12.00	$10,968.00	$1,645.20	$9,322.80	$2,742.00	$6,580.80	01-12-2014	12	December	2014
683	Government	France	Paseo	High	$293.00	$10.00	$20.00	$5,860.00	$879.00	$4,981.00	$2,930.00	$2,051.00	01-12-2014	12	December	2014
684	Channel Partners	Mexico	Velo	High	$500.00	$120.00	$12.00	$6,000.00	$900.00	$5,100.00	$1,500.00	$3,600.00	01-03-2014	3	March	2014
685	Midmarket	France	Velo	High	$2,826.00	$120.00	$15.00	$42,390.00	$6,358.50	$36,031.50	$28,260.00	$7,771.50	01-05-2014	5	May	2014
686	Enterprise	France	Velo	High	$663.00	$120.00	$125.00	$82,875.00	$12,431.25	$70,443.75	$79,560.00	$(9,116.25)	01-09-2014	9	September	2014
687	Small Business	United States of America	Velo	High	$2,574.00	$120.00	$300.00	$7,72,200.00	$1,15,830.00	$6,56,370.00	$6,43,500.00	$12,870.00	01-11-2013	11	November	2013
688	Enterprise	United States of America	Velo	High	$2,438.00	$120.00	$125.00	$3,04,750.00	$45,712.50	$2,59,037.50	$2,92,560.00	$(33,522.50)	01-12-2013	12	December	2013
689	Channel Partners	United States of America	Velo	High	$914.00	$120.00	$12.00	$10,968.00	$1,645.20	$9,322.80	$2,742.00	$6,580.80	01-12-2014	12	December	2014
690	Government	Canada	VTT	High	$865.50	$250.00	$20.00	$17,310.00	$2,596.50	$14,713.50	$8,655.00	$6,058.50	01-07-2014	7	July	2014
691	Midmarket	Germany	VTT	High	$492.00	$250.00	$15.00	$7,380.00	$1,107.00	$6,273.00	$4,920.00	$1,353.00	01-07-2014	7	July	2014
692	Government	United States of America	VTT	High	$267.00	$250.00	$20.00	$5,340.00	$801.00	$4,539.00	$2,670.00	$1,869.00	01-10-2013	10	October	2013
693	Midmarket	Germany	VTT	High	$1,175.00	$250.00	$15.00	$17,625.00	$2,643.75	$14,981.25	$11,750.00	$3,231.25	01-10-2014	10	October	2014
694	Enterprise	Canada	VTT	High	$2,954.00	$250.00	$125.00	$3,69,250.00	$55,387.50	$3,13,862.50	$3,54,480.00	$(40,617.50)	01-11-2013	11	November	2013
695	Enterprise	Germany	VTT	High	$552.00	$250.00	$125.00	$69,000.00	$10,350.00	$58,650.00	$66,240.00	$(7,590.00)	01-11-2014	11	November	2014
696	Government	France	VTT	High	$293.00	$250.00	$20.00	$5,860.00	$879.00	$4,981.00	$2,930.00	$2,051.00	01-12-2014	12	December	2014
697	Small Business	France	Amarilla	High	$2,475.00	$260.00	$300.00	$7,42,500.00	$1,11,375.00	$6,31,125.00	$6,18,750.00	$12,375.00	01-03-2014	3	March	2014
698	Small Business	Mexico	Amarilla	High	$546.00	$260.00	$300.00	$1,63,800.00	$24,570.00	$1,39,230.00	$1,36,500.00	$2,730.00	01-10-2014	10	October	2014
699	Government	Mexico	Montana	High	$1,368.00	$5.00	$7.00	$9,576.00	$1,436.40	$8,139.60	$6,840.00	$1,299.60	01-02-2014	2	February	2014
700	Government	Canada	Paseo	High	$723.00	$10.00	$7.00	$5,061.00	$759.15	$4,301.85	$3,615.00	$686.85	01-04-2014	4	April	2014
701	Channel Partners	United States of America	VTT	High	$1,806.00	$250.00	$12.00	$21,672.00	$3,250.80	$18,421.20	$5,418.00	$13,003.20	01-05-2014	5	May	2014
702																

This data set has total of 700 observations and 16 dimensions.

Data Processing and Pipeline

Cleaning Steps: Procedures for handling missing values, duplicate records, and irrelevant data points.

Imputation Techniques: Methods used to estimate and fill in missing data.

Transformation Processes: Normalization (Z score or T score), scaling, or any other transformation applied to the data to prepare it for analysis.

Outlier Detection and Handling: Strategies to identify and manage outliers that could skew the analysis. (Box-plot)

Data Stylized Facts

Distributional Analysis: Summary of the key characteristics of the data distributions, including normality assessments and distribution shapes.

Clustering Observations: Findings from any clustering techniques used to identify natural groupings within the data.

Dimensionality Reduction: Techniques such as PCA (Principal Component Analysis) used to simplify the dataset while retaining the most informative features.

Model Selection

Feature Selection Requirements: Criteria used to determine which features are included in model development.

Classification/Regression Approaches: Overview of the methodologies used for predictive modelling, including the rationale for selecting certain models.

Reference/Baseline Model: Details of a baseline model used for comparative performance assessment.

Software Packages, Applications, and Tools

Programming Languages: R and RStudio.

Libraries and Packages: dplyr, tidyr, quantmod, ggplot2, forecast, shiny, rmarkdown

Tools: SQL databases

Project Management and Source Control: GitHub

Data Visualization tool: Tableau

Reference Resources

- Financial Statement Data Sets - SEC.gov: https://www.sec.gov/dera/data/financial-statement-data-sets
- US securities Financial Statements Data Sets: https://www.sec.gov/files/aqfs.pdf
- Business: Financial Databases: Referencing Financial Data - Library Guide: https://libguides.mdx.ac.uk/c.php?g=322130&p=4861962
- How do I reference financial data - Cranfield University Blogs: https://blogs.cranfield.ac.uk/library/financials-apa7/

Supplemental Resources

- Financial Data from an Online Database Example: In-text citation: During 2018, Vodafone Group plc showed increased operating revenues (S&P Capital IQ, 2019).
- Example: In-text citation: The company's profits expanded (M&S plc, 2015) ..... Reference list: M&S plc. (2015) Annual report 2015. Available at: http://annualreport2015.marksandspencer.com/ (Accessed: 8 January 2016).
- Test your pre-processing skills with this dataset!! - Company Financials Dataset - Kaggle: https://www.kaggle.com/datasets/atharvaarya25/financials