॥ सा विद्या या विमुक्तये ॥

भारतीय प्रौद्योगिकी संस्थान धारवाड

Indian Institute of Technology Dharwad

Project Report

Panel Data Analysis

Ashwin Waghmare
Department of Computer Science
210010060

Project Manager
Prashant Bhoske

2 May - 31 July
2023

Acknowledgements

I would like to express my sincere gratitude to TATA Technologies Ltd. for granting me the opportunity to intern with your esteemed organization. I am truly honored to have been selected for this internship to contribute my skills and learn from the talented professionals at your company.

I am thankful to my supervisor, Mr Prashant Bhoske for presenting me with a unique chance for me to gain practical insights into the industry, refine my skills, and grow both personally and professionally.

I would also like to extend my appreciation to my mentors, Mr Siddhesh Deo and Mr Harshwardhan Jadhav who guided me throughout this internship. Your guidance and expertise undoubtedly played a crucial role in shaping my understanding of the industry and enhancing my skill set.

I am grateful to Mr Mahendra Waghmare for his unwavering support throughout. Your support has made the difference, and I am truly fortunate to have your helping hand.

Lastly, I would like to extend my thanks to the HR and IT departments of Tata Technologies Ltd. for ensuring a smooth onboarding process.

Contents

1 Industry Overview

Tata Technologies Ltd. was founded in 1989 as a subsidiary of Tata Motors, a prominent Indian automotive manufacturing company. Tata Technologies provides a range of services aimed at helping companies design, engineer, and manufacture products efficiently. Some of its key offerings include:

- **Engineering Services:** Tata Technologies offers a wide array of engineering solutions, including mechanical design, virtual simulation, product development, and testing. These services are used across industries to enhance product quality and reduce time-to-market.
- **Manufacturing Solutions:** Tata Technologies offers manufacturing process optimization, tooling design, and manufacturing support services to help companies improve their production processes and enhance efficiency.
- **Consulting Services:** Tata Technologies provides consulting services to address various business challenges, including supply chain optimization, cost reduction, and innovation strategy.
- **Automotive:** Tata Technologies has deep roots in the automotive industry, providing design, engineering, and manufacturing support to automotive manufacturers and suppliers worldwide.
- **Aerospace and Defense:** The company offers services to aerospace and defense companies, including aircraft design, simulation, and manufacturing solutions.

2 Executive Summary

Researching Panel Data Analysis

With Safety Incidents Forecasting in mind the purpose was to research useful concepts that can be implemented to better the existing model. Panel Data analysis is aimed to understand the relationships between variables, account for individual-specific effects, and uncover potential causal links. The panel dataset encompassed information from several factories over a time period.

3 Time Series Forecasting

This section provides an overview of time series forecasting techniques and their application in predicting future values based on historical data. It covers the fundamentals of time series analysis, discusses various forecasting methods, and highlights their strengths and limitations. The report also includes a case study demonstrating the practical implementation of time series forecasting using real-world data.

3.1 Stationary Series

To build a time series model the given data series must be stationary series (If not, make it!). Stationary series means that the statistical properties of a process generating a time series do not change over time. It does not mean that the series does not change over time, just that the way it changes does not itself change over time. A stationary time series exhibits constant mean, variance, and autocovariance over time.

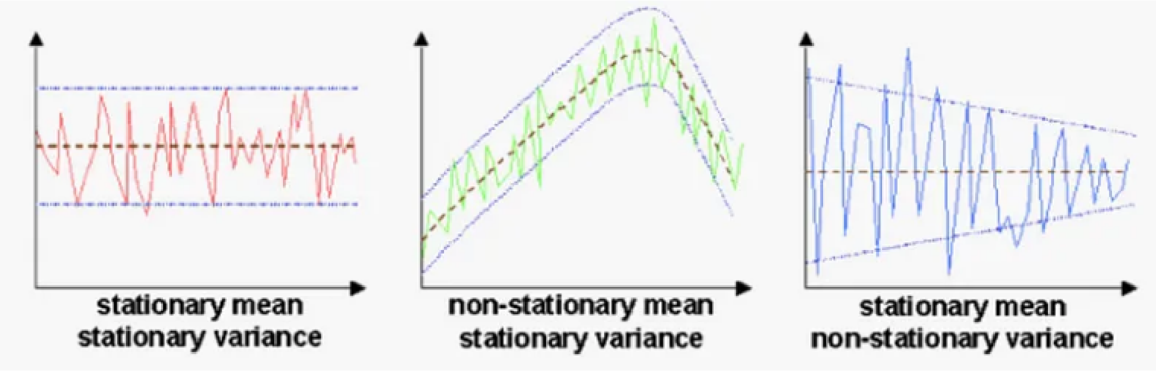


Figure 1: Comparing graphs of Series

3.1.1 ADF test

One of the most widely used statistical tests is the Augmented Dickey-Fuller test. It can be used to determine whether or not a series has a unit root, and thus whether or not the series is stationary. This test's null and alternate hypotheses are:
Null test: The series has a unit root.
Alternate Hypothesis: The series has no unit root.
If the null hypothesis is not rejected, the series is said to be non-stationary. The series can be linear or difference stationary as a result of this. The series becomes stationary if both the mean and standard deviation are flat lines (constant mean and constant variance).

3.2 ARIMA model

Now, our challenge is to convert nonstationary series to stationary. The ARIMA model is the most widely used one, to perform this task. Its component are as follows:

1. **Autoregressive (AR) Component**
   The autoregressive component represents the dependency of the current value on

previous values within the series. It assumes that the current value can be linearly predicted based on its own lagged values. The order of the autoregressive component, denoted as AR(p), determines the number of lagged values used for prediction.

2. **Integrated (I) Component**
   The integrated component involves differencing the time series data to achieve stationarity. Differencing removes trends or seasonality, making the series suitable for ARIMA modeling. The order of differencing, denoted as I(d), specifies the number of times the differencing operation is applied to achieve stationarity.

3. **Moving Average (MA) Component**
   The moving average component considers the dependency of the current value on past forecast errors. It assumes that the current value can be explained by a linear combination of the previous errors. The order of the moving average component, denoted as MA(q), determines the number of previous errors considered in the model.

Each technique has its own assumptions, strengths, and limitations. The choice of a specific model depends on the characteristics of the data and the desired forecasting horizon.

4 Feature Engineering

Feature engineering is the process of transforming raw data into features that better represent the underlying data. It is a crucial process in model building as the quality of features that we feed to the model is the primary limiting factor on the performance of that model. Let us now look at some techniques associated with feature engineering:

4.1 Techniques of Feature Engineering

4.1.1 Imputation

Feature engineering deals with inappropriate data, missing values, human interruption, general errors, insufficient data sources, etc. Missing values within the dataset highly affect the performance of the algorithm. Imputation is used to tackle this problem. **Imputation is responsible for handling irregularities within the dataset.**

For example, removing the missing values from the complete row or complete column by a huge percentage of missing values. But at the same time, to maintain the data size, it is required to impute the missing data, which can be done as:

- For numerical data imputation, a default value can be imputed in a column, and missing values can be filled with means or medians of the columns.

- For categorical data imputation, missing values can be interchanged with the maximum occurred value in a column.

4.1.2 Outliers

Outliers are the deviated values or data points that are observed too away from other data points in such a way that they badly affect the performance of the model. Outliers can be handled with this feature engineering technique. This technique first identifies the outliers and then remove them out.

Standard deviation can be used to identify the outliers. For example, each value within a space has a definite to an average distance, but if a value is greater distant than a certain value, it can be considered as an outlier. **Z-score** can also be used to detect outliers.

4.1.3 Log Transform

Log transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

4.1.4 Binning

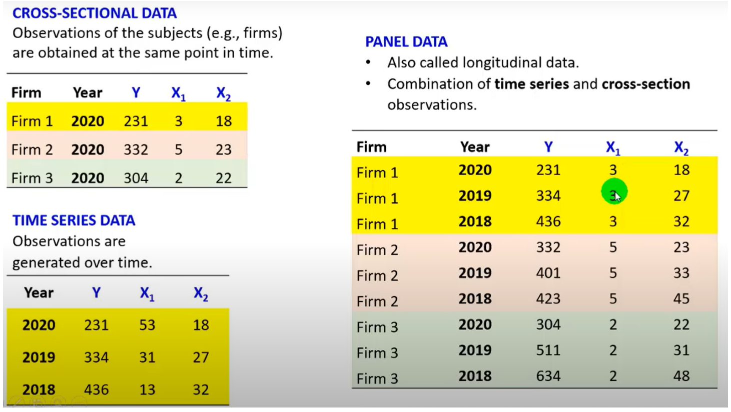
Binning is used to normalize the noisy data. This process involves segmenting different features into bins. As the noisy data is normalized, the model does not face overfitting.

4.1.5 Feature Split

Feature split is the process of splitting features intimately into two or more parts and performing to make new features. This technique helps the algorithms to better understand and learn the patterns in the dataset.

5 Panel Data Analysis

Panel data analysis is a statistical method used to analyze data collected over time and across different entities or individuals. The data used, Panel data, is a combination of Time series data and Cross-sectional data. The following illustration better explains it.



CROSS-SECTIONAL DATA

Observations of the subjects (e.g., firms) are obtained at the same point in time.

Firm	Year	Y	X_1	X_2
Firm 1	**2020**	231	3	18
Firm 2	**2020**	332	5	23
Firm 3	**2020**	304	2	22

TIME SERIES DATA

Observations are generated over time.

Year	Y	X_1	X_2
2020	231	53	18
2019	334	31	27
2018	436	13	32

PANEL DATA

- Also called longitudinal data.
- Combination of **time series** and **cross-section** observations.

Firm	Year	Y	X_1	X_2
Firm 1	**2020**	231	3	18
Firm 1	**2019**	334	3	27
Firm 1	**2018**	436	3	32
Firm 2	**2020**	332	5	23
Firm 2	**2019**	401	5	33
Firm 2	**2018**	423	5	45
Firm 3	**2020**	304	2	22
Firm 3	**2019**	511	2	31
Firm 3	**2018**	634	2	48

Figure 2: Comparing Dataset of Models

There are several models available and choosing the appropriate panel data model depends on the research question and the underlying assumptions of the data. If the individual-specific effects are uncorrelated with the independent variables, the Pooled OLS model can be used. However, if there is potential correlation between the individual-specific effects and the error term, the Random Effects model is more appropriate. On the other hand, if the focus is on controlling for individual-specific effects and capturing within-individual variations, the Fixed Effects model should be employed. The models are further elaborated below.

5.1 Pooled OLS model

The Pooled OLS model is the simplest approach in panel data analysis, which treats the data as if it were cross-sectional and ignores the individual-specific or time-specific effects. It pools all observations together and estimates a single regression equation using the entire dataset. However, this model assumes that the individual-specific effects are uncorrelated with the independent variables, which may lead to biased and inefficient estimates. The formula for a pooled OLS model can be written as follows:

$$Y = X\beta + \epsilon \tag{1}$$

where:

- Y is the dependent variable, typically a vector of observed values.
- X is the matrix of independent variables, including a constant term (intercept) and explanatory variables.
- β is the vector of coefficients (parameters) to be estimated.
- ϵ is the vector of error terms, which captures the unobserved factors affecting the dependent variable.

5.2 Random Effects Model

The Random Effects model accounts for both the individual-specific effects and the time-specific effects. It assumes that the individual-specific effects are uncorrelated with the independent variables but allows for correlation between the individual-specific effects and the error term. This model captures the heterogeneity across individuals, assuming that the individual-specific effects are random variables drawn from a larger population. The random effects model can be formulated as follows:

$$Y_{it} = X_{it}\beta + \alpha_i + \epsilon_{it} \tag{2}$$

where:

- Y_{it} is the dependent variable for unit i at time t.
- X_{it} is a vector of independent variables for unit i at time t.

- β is a vector of coefficients (parameters) to be estimated, representing the effect of the independent variables.
- α_i s the individual-specific effect or random effect for unit i. It captures the unobserved heterogeneity that is constant over time for each unit.
- ϵ_{it} s the error term, representing the random component of the model for unit i at time t.

5.3 Fixed Effects Model

The Fixed Effects model controls for the individual-specific effects by including a separate intercept for each individual in the regression equation. It assumes that the individual-specific effects are correlated with the independent variables but uncorrelated with the error term. By controlling for individual-specific effects, the Fixed Effects model allows for within-individual comparisons over time. Estimation of the Fixed Effects model can be done using within-group transformations or the Least Squares Dummy Variable (LSDV) approach.

6 Safety Incidents Forecasting

Safety Incidents forecasting is an application of machine learning algorithms stated in above sections. Safety incidents occurring in a factory like, thumbs being injured by hammer, are recorded over time and with the help of this application the number of incidents occurring on particular day can be predicted so as to cation the factory workers.

The following is a flowchart depicting the structure of this application.

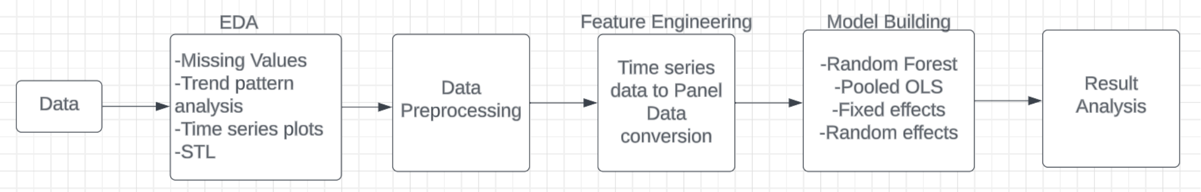


Figure 3: Working of model

- Data is collected from factories with proper date-time and description specifics.
- This data undergoes Exploratory Data Analysis wherein it is checked for missing values, its trend patterns are analysed through time series plots.
- This time series data is then converted into panel data.
- Once we have a suitable dataset, we select an optimum model.
- The obtained result is then interpreted through graphs and can be used to caution workers accordingly.

This application has been deployed and is in use by various divisions of the factory. It has greatly reduced the hazard imposed to the workers.

7 Conclusion

In conclusion, the panel data analysis unveiled significant insights into the relationships between variables and their effects on the outcome. The findings provide valuable guidance for safety officers in factories emphasizing the need for continued investigation into the identified factors.

8 References

- Artificial Intelligence a Modern Approach by Stuart Russell and Peter Norwig
- A First Course in Artificial Intelligence by Deepak Khemani
- Linear Regression by Jia Bin Huang, Virginia Tech
- Hyndman, R. J., Athanasopoulos, G. (2018). Forecasting: principles and practice (2nd ed.). OTexts.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. (2015). Time series analysis: forecasting and control (5th ed.). Wiley.
- Zhang, G., Patuwo, B. E., Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35-62.
- Econometric Analysis of Panel Data by Badi H. Baltagi
- https://www.geeksforgeeks.org/supervised-unsupervised-learning/
- https://www.youtube.com/watch?v=W01tIRP_Rqs&ab_channel=IBMTechnology
- https://www.javatpoint.com/data-preprocessing-machine-learning
- https://www.javatpoint.com/machine-learning-algorithms
- https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
- https://www.geeksforgeeks.org/decision-tree/
- https://www.javatpoint.com/machine-learning-models
- https://deepchecks.com/how-to-test-machine-learning-models/
- https://www.javatpoint.com/performance-metrics-in-machine-learning
- https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide
- https://www.javatpoint.com/feature-engineering-for-machine-learning
- https://serokell.io/blog/feature-engineering-for-machine-learning
- https://towardsdatascience.com/a-guide-to-panel-data-regression-theoretics