BUSINESS DATA MANAGEMENT - PROJECT
FINAL SUBMISSION

SEPTEMBER TERM - CYCLE III

OPTIMIZING CASH FLOW AND STAGGERED CREDIT SALES

SYSTEM FOR A B2B TEXTILE COMPANY

SUBMITTED BY
NAME : Shriram G
ROLL NO.: 21f2000668

B.S. DATA SCIENCE AND APPLICATIONS

2021-2024

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS

EXECUTIVE SUMMARY:

This business data management project was conducted for a B2B textile company that specializes in the purchase of yarn and sale of poplin cloth. The company operates in a highly competitive market and faces challenges in optimizing its purchasing and sales processes while minimizing bad debtor occurrences.

To address these challenges, the company collected and analyzed data related to its yarn purchasing, sales, and bad debtor occurrences. The project aimed to improve the company's understanding of the relationship between various input data and bad debtor occurrences, as well as identify factors influencing yarn purchase and sales.

To achieve these objectives, the project utilized state space analysis and linear regression modeling to analyze the data and identify trends and patterns. The linear regression analysis identified factors influencing yarn purchase and sales, and the regression model was used to model state variables for the state space method of analysis.

The state space analysis was done for the ability to identify the most significant factors influencing bad debtor occurrences, understand the dynamics of the purchasing and sales processes, and make predictions about future occurrences.

By gaining a deeper understanding of these factors, the company would be able to optimize its operations and minimize the risk of bad debtors, resulting in improved efficiency in cash flow and profitability.

The various pitfalls of such analysis methods, the future scope of analysis, the justification for such analysis methods and the shortcomings of the current available methods and data are also discussed.

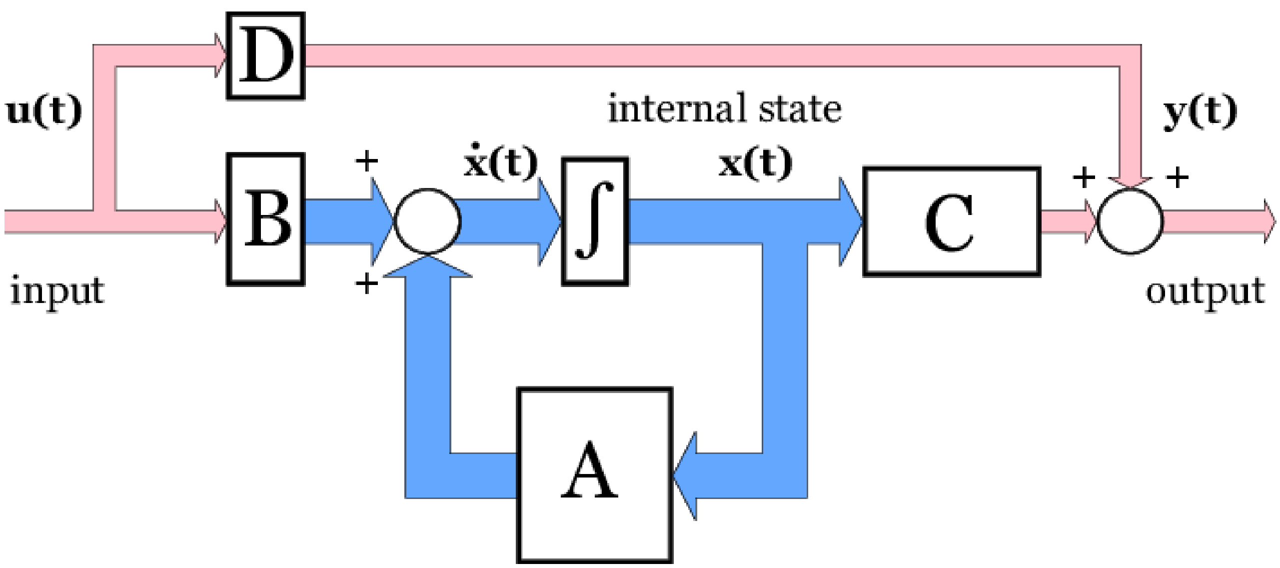
The final results and findings and the corresponding conclusions and recommendations for the business to meet the objectives of resolving their current business issues were determined and are presented in this report.

ANALYSIS PROCESS AND METHOD:

To implement any control system design via state space analysis, data for the following equations is required.

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t)$$

In this equation, x(t) is the state vector. The state vector describes state variables of our system, which in our case, is the business of concern. x'(t) is the derivative of the state vector. A, B, C and D are the System, Input, Output and Feedforward matrices respectively.



State Space System - Block Diagram Representation

They are of dimensions that depend upon the number of state variables we model our system on. In our case, the number of state variables is three. Our state variables at given time instance 't', from the equation are the following in respective order in the state vector used in the training while developing our model:

1. Yarn Price
2. Sales in Metres of Cloth
3. Losses Accumulated

The symbol u(t) represents the control input vector, which will be the total amount of yarn chosen to be purchased at the given time. Our aim is to use data to develop a state space model with matrices A, B, C and D so that we may perform predictive analytics at any given time instance with our knowledge of the state variables of the business.

According to the Theory of Control Systems/Control System Engineering, it is widely known that for non-purely-deterministic systems, it is not mathematically possible to determine the physical or observable meaning of state variables. If we impose physical meaning on state variables and use them to calculate the state space system, it may not be possible to qualitatively determine the parameters of the state matrices A, B, C, and D.

Our business model focuses on developing this system such that the business person might utilize it with knowledge of state variables to predict the output of the system. Hence, we choose to constrain the state variables to the ones mentioned above (yarn price, etc.) and quantitatively determine the estimated values of the elements of the A, B, C and D matrices.

This may be done via initializing these matrices with random entries and training the system parameters with the data available. The data available for efficient training however is very highly limited. Transactions of the business are mass transactions of lakhs of rupees and training the system to find the optimal set of parameters becomes very difficult with such such less number of datapoints with such high transactional value.

Hence, we instead develop a continuous equation representation model using regression analysis for predicting the rate of yarn, the sales and losses of the system as a moderate smoothing measure for the curve instead of random and erratic discrete data points.

Hence, the regression model's estimated price, sales and losses which is a gradual function instead of the haphazard curve we may get if we only plot the mass transfer of cash at each data point, we may get a smooth system that gives an idea of the long term and gradual effect of a control input and state variation at any given time instance to an arbitrary precision. In this case, preferably a day.

Therefore in our training approach, a regression model that is representative of the state of the system at a time will be used as a substitute of the state vector for all 365 days of the year instead of the 30-50 samples we have of the data.

Hence, during training, the regression model's predicted yarn price, sales and losses at time instant 't' will be the state variables at time instant 't'. The predicted buying pattern of the business person to meet demands is in the kilograms of yarn purchased and that will be the control input.

However, for a proper and stable system, for the given data, we must have a stable frequency response for an impulse signal, which is a control input of 'one'. To test this, we must employ a trial and error method of estimating A, B, C and D matrices. We initialize these matrices' values randomly via a gaussian distribution. Each matrix is initialized with a different combination of means for the normal(gaussian) method of random initialization.

To validate the correctness of these matrices and their description of the system, there are two conditions. One is to check the stability of the impulse response of the system. The business that is being analyzed has had steady profits and moderate losses only.

It is clear from the appearance of the balance sheet. The time response might have a haphazard appearance, but it follows a pattern to yield such a steady profit and loss output year after year for this business.

Hence, the frequency response of the system must be a steady set of values. The absolute value of the state vector must be constrained within a reasonable value, for this business which has a crore or more in turnover each year.

For the final step of resolving the issue of estimating parameters and beginning training of the system, we need the derivative state-vector, which we also cannot determine effectively simply using the data and formulas for slope/gradients of a line.

We cannot model the data using differential equations before we determine the state vector derivative as differential equation analysis would be too variable for such data. Therefore, only a poorer numerical approximation from the equation is possible to be arrived at, with the validation measures stated above to assure a general idea of the system's nature.

Subsequently, a formula for credit scores to implement a staggered credit sale system compatible with the state space system is implemented to streamline the business process and solve the business' core issues in bad debtor sales and management.

All analysis processes were conducted with the Python programming language and necessary external library packages, namely - Control Systems Library, Scikit-Learn, Numpy, Pandas and Matplotlib.

RESULTS AND FINDINGS:

1. STATE SPACE ANALYSIS:

Upon implementation of our analysis processes, the final arrived state space equation is given below. The order is read from top-left to bottom-right - A, B, C, D.

A = Top-Left. B = Top-Right. C = Bottom-Left. D = Bottom-Right

As mentioned in the analysis process, the matrices were initialized with a trial and error pattern on a Gaussian/Normal distribution of values. The mean of the Gaussian of each of the matrices is 75, 10, 200 and 50 for A, B, C and D respectively.

This was initialized and experimented on with the same means above with using Numpy's pseudo-random generators. The final arrived value for the above means with a random seed is given below.

$$
\left(\begin{array}{ccc|ccc}
75.3 & 75.9 & 76.1 & 12.5 & 10.4 & 9.46 \\
74.1 & 74.7 & 78 & 9.63 & 9.32 & 9.33 \\
76.2 & 74.2 & 73.6 & 9.75 & 8.72 & 8.95 \\
\hline
201 & 197 & 199 & 50.1 & 48.7 & 50.9 \\
197 & 200 & 200 & 49.3 & 50.1 & 50.1 \\
198 & 200 & 199 & 51 & 48.6 & 50.2
\end{array}\right)
$$

Gaussian Distributed State Space System Matrices

The vector of state variable derivatives is given below :

```
[[2472166.30721101]
 [2528413.68045938]
 [2388133.36730192]]
```

State Vector Derivative - 3 x 1 Matrix

We notice that this also delivers an apt, well constrained derivative across a year's worth of variable data, for a business with crores in turnover.

Upon having the necessary matrices above, we generate the frequency response of the system.
The frequency response may be tested with relevant test signals. Various test signals exist, but the most fundamental test signal is the impulse signal function. In the discrete case, an impulse response consists of only a test input of unit 1.

The magnitude response is a spectrum of frequencies and their respective magnitude output of the system at those frequencies. A Bode Plot is not generatable via python for MIMO systems (multiple-input - multiple-output).

However, the impulse response matrices are generated and attached below.

```
[[[24.751649  ]
  [25.19626607]
  [26.8803163 ]]

 [[17.25861997]
  [23.11780816]
  [25.33327167]]

 [[18.92571673]
  [21.73839999]
  [25.56195673]]]
```

Magnitude Response Matrix For Impulse Signal Input

We notice that the magnitude response is more or less clustered around the mean. The cycles that usually occur in time-series data are identified here in a pattern for the input variations of the system. Hence the system has a uniform magnitude response.

The phase response of the system shows the correlation of signal/input and output signal with the reference zero signal’s phase. In the context of a business system, it may be inferred that this as a change in system response correlation between the relevant inputs and outputs.

This business system is found to be highly correlated in all its inputs and state variables which is apparent from the negligible phase values of the phase plot. It should be noted that the phase response value is not in the typical range of 0 to 1 and it is in the units of radians.

It denotes the radians in decimal format. 1 radian is approximately 57 degrees and Pi radians, where Pi is the infamous mathematical constant that is the ratio of the circumference of a circle to its diameter, is 180 degrees in phase angle. A complete 180 degree phase value is equivalent to negative correlation. A 90 degree response, is half-pi radians is equivalent to absolutely no correlation. Here, we get a near 0 degree response which denotes high positive correlation.

It is a highly interdependent system where one factor may affect any other factor. Since we have constrained the state variables for our convenience to physically significant or observable variables, our system matrices remain unobservable Gaussian distribution parameters.

```
[[[-0.03716289]
  [-0.02699733]
  [-0.00984851]]

 [[ 0.22483765]
  [ 0.08790939]
  [ 0.0134391 ]]

 [[ 0.27330537]
  [ 0.12063203]
  [ 0.01715614]]]
```

Phase Response Matrix For Impulse Signal Input

They may only be treated as weights of a curve and the observable meaning of those parameters, cannot be inferred. If we wish to identify system parameters, their correlation etc., like we have for the state variable parameters, we must randomly initialize state variables and redo the analysis.

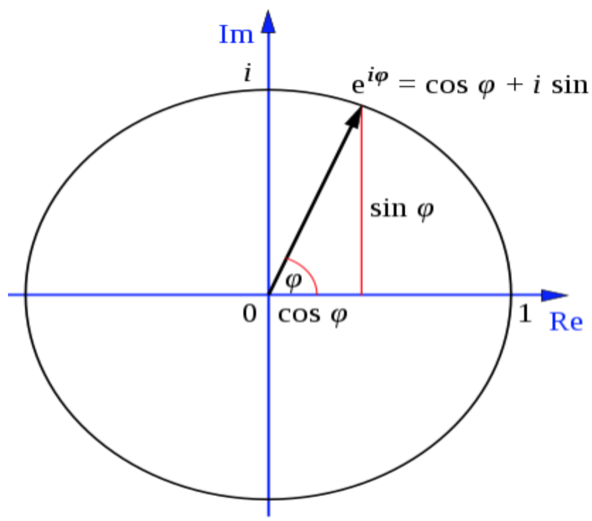
But for the purpose of this problem, we only have state variables available as business data to train and identify the system to conduct analysis. Hence we are restricted mathematically to this approach.

Thus the magnitude plot and phase plot consist of the frequency response of the system. It is to be noted that these responses are in the frequency domain, also known as the Fourier Domain or Fourier-Transform Domain. This naive transformation yields real and complex values and the norm of the complex values at respective frequencies consists of the phase plot.

The implementation of each formula is abstracted from us via the wrapper library functions provided by Python libraries. The general formula for calculating magnitude plot from the complex domain values is given below.

$$
|z| = \sqrt{zz^*} = \sqrt{(x+iy)(x-iy)}
$$
$$
= \sqrt{x^2 + y^2}
$$

Magnitude of Complex Number : Re = Real part and Im = Imaginary part



Phase of a Complex Number (here denoted by Phi)

This is an Argand Plane (Complex Plane) visualization of the phase and magnitude.
This helps in visualizing the above described measures of correlation and meaning of the magnitude response value.

2. REGRESSION MODELLING:

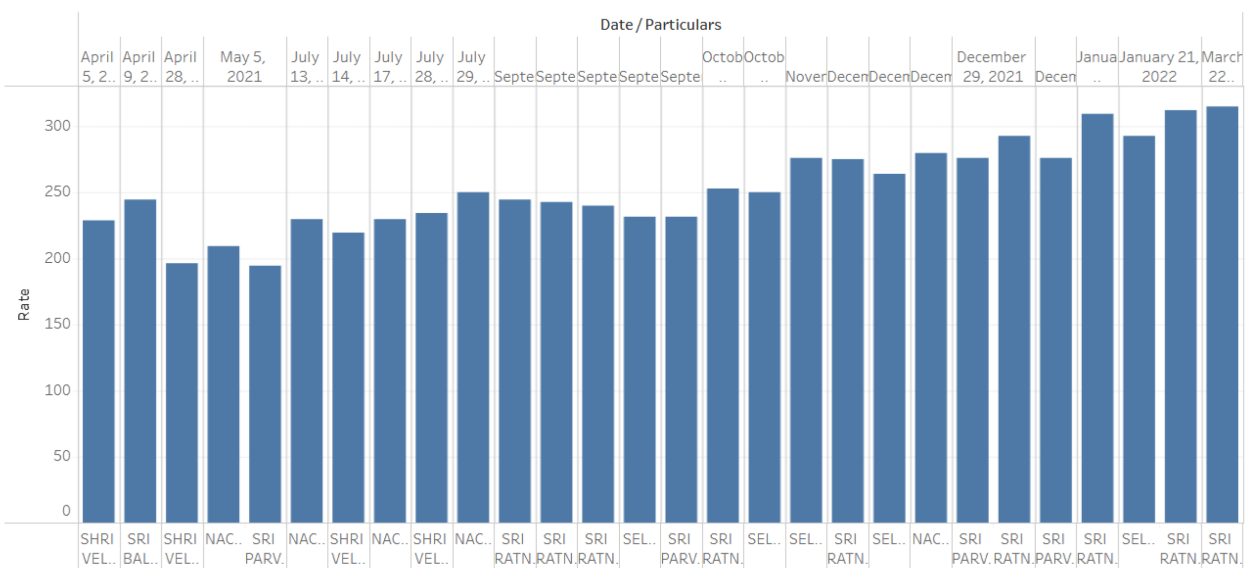
In the state space analysis done above, instead of using individual data points, we used the predicted value of the regression model at each instance throughout the year to predict the output of the system, for all the associated state variables and the control input. The regression model is detailed below.

Multiple linear regression was performed to predict the values in this model for each variable. This means multiple independent variables were used to predict the dependent variable such as yarn price, sales and loss.

For the purpose of fitting Date data into the model, the date values have been formatted into the proleptic Gregorian Ordinal format of the date. This means a date of 2021-08-03 (YYYY-MM-DD) would be denoted and used for all purposes of calculations as 738005 as this is standard practice while using long-term time based data in Scikit-Learn and Pandas

Linear Regression is the preferred method of modelling as the preliminary survey of the data clearly indicated that while the system's total functionality is complex as it is, pairs of variables needed to replace state vector data have a high degree of linearity in their trendlines.

A tableau viz is attached to visualize the trend of the data.



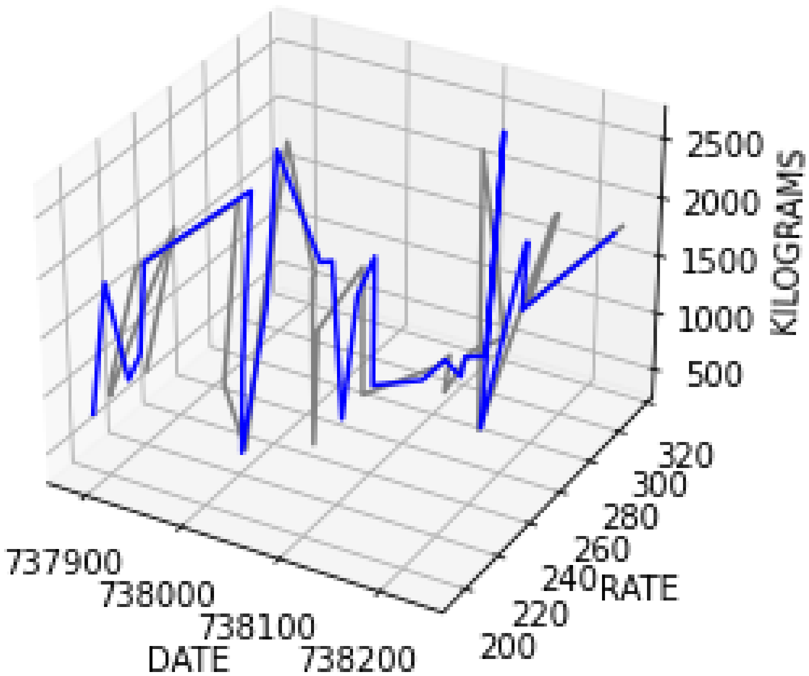
Date Of Purchase vs Rate of Yarn

3. YARN PRICE FORECASTING:

The yarn price was found to follow a good linear model as expected in the initial findings upon scrutinizing the data. The linear regression equation that was used to determine the yarn price prediction curve is, Yarn Price = 0.2982144 * Date + 0.00514733 * Kilograms - 219852.35543

We see it is a moderate slope and high intercept, which is obvious from the gradual trend observed from the data and the high volume transactions and mile variation in price change for unit volume of purchase

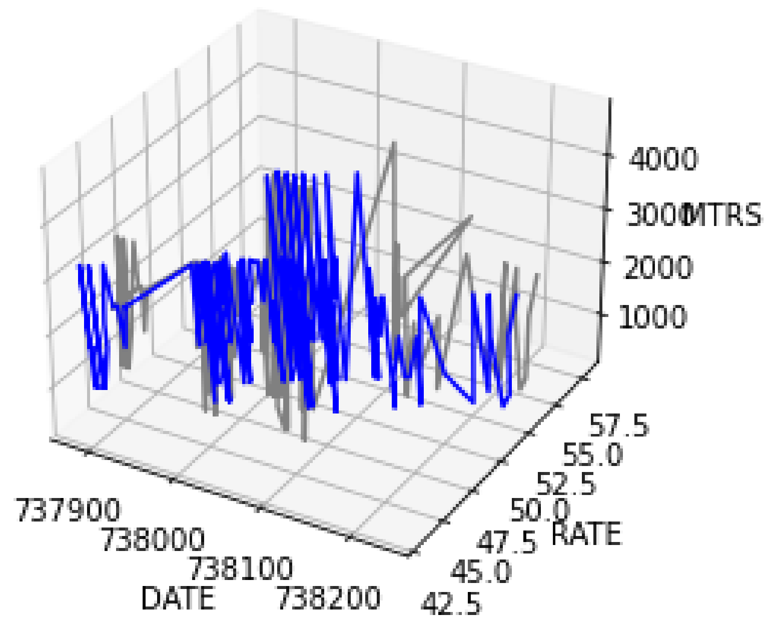
'Grey' colour denotes original data line and 'Blue' colour denotes the line drawn for inputs in regression equation



Linear Regression - Yarn Price

Similarly, for poplin sales,

Sales = 0.02705822 * Date - 0.00031816 * Meters - 19921.505265947093

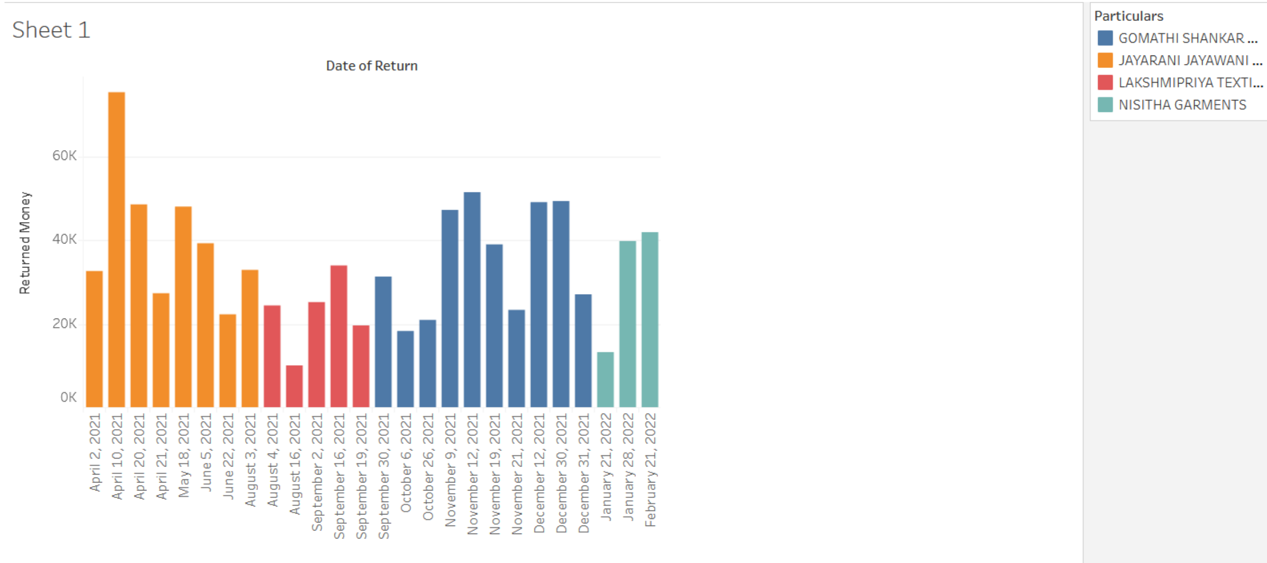


Linear Regression - Sales

We can clearly observe that the choice of using linear regression to extend our data points for training the state space system was aptly justified considering the regression line is able to approximate the original data pattern considerably well. Hence the use of regression analysis on this data gives us deeper insight into the nature of the firm's state variables and its patterns.

To validate this with metrics rather than visual interpretation, the R2 score is used. The R2 score is the coefficient of determination of the regression model. Our models follow on average an R2 score of 0.82 which is good for such a dataset.

This validates our analysis algorithms and methods for this dataset. Once the models are built, the debtor data is analyzed. The state space system and debtor data must work in tandem to establish a staggered credit sales system for this business



Debtor Return Data

This chart illustrates the amount of money returned by debtors who defaulted. Gomathi Shankar The data is presented with the company name in colors for ease of referring to the chart.

Here we notice that there is no linear relationship between the loss incurred by the company and the amount borrowed. On comparing and contrasting the return data chart, we can see that while company ‘Blue’ had the highest consistent return pattern, it has incurred the highest loss while a company such as ‘Sea Blue’ with no consistent return pattern made less of a loss.

COMPANY	RETURN	PURCHASE	LOSS	%LOSS
Gomathi Shankar	357620.14	578222	220601.86	39
Jayarani Jayawani	326555.98	411313	84757.02	21
Lakshmipriya Textiless	113502.54	147411	33908.46	24
Nisitha Garments	94877.32	141363	46485.68	33

4. STAGGERED CREDIT:

Here we notice that there is no linear relationship between the loss incurred by the company and the amount borrowed. On comparing and contrasting the return data chart, we can see that while company ‘Gomathi Shankar’ had the highest consistent return pattern, it has incurred the highest loss while a company such as ‘Nisitha Garments’ with no consistent return pattern made less of a loss.

It is impossible to judge the company’s return pattern to judge their loss or borrowing capabilities in the subsequent months. Hence, this data must also be fed into the state space equation and that was what was done in the choice of state variables. Companies are given a default timeline of 2-4 months in the process of making sales and returning the money in a given amount of time.

When they do not, the amount of yarn purchased is to be reduced as the cloth returned by the company would be enough to compensate for the demand. The returned cloth can also be treated as its respective cash value for the purpose of developing the model which is what was done in the debtor regression line which was the third state variable. This accounts for making investments for the subsequent quarter in purchase of yarn for manufacturing and distribution.

Credit Score : Hence a credit score of the business owner’s interest may be assigned and the company’s sorted accordingly in a hierarchical manner. The model to find the control input (total yarn purchase) that can be spared for a healthy state of the business without loss and in expectation of a profit is now complete and the business owner may supply the company’s of his interest according to the credit score rating available.

Since the state vector derivative is a significant percentage of total turnover, the model will be tolerant of errors for at least 2 months in advance, provided there is not a huge bad debt being incurred. An appropriately weighted sample credit score formula proposed would be as follows (purely for debtors) :

Score = Time(months)/ 2 or 4(months) + Fraction of Loss

Since 4 months is usually the deadline in the business cycle to settle all payments, that would amount to a 1 on the deadline and be higher if the due date is exceeded and a loss is incurred.
In this model, the lower the credit score, the better the company to trust.
The debtor's credit score is tabulated below using the formula above and the portfolio of bad debtors is formulated in order and a staggered credit system is implemented.

DEBTOR	**SCORE**
Gomathi Shankar	3/4 + 0.39 = 1.14
Jayarani Jayawani	4/4 + 0.21 = 1.21
Lakshmipriya Textiles	2/2 + 0.24 = 1.24
Nisitha Garments	2/2 + 0.33 = 1.33

We can see how smaller buyers with a considerable percentage of losses are penalized almost equally as larger buyers who have returned a much higher percentage of the cash in return.
While this may not be an entirely fair credit sale system, it can be further optimized by adjusting the formula accordingly:

Score = Time To Return (days)/60 or 120 days + Fraction of Loss
Days or weeks might give a better running credit score that reflects the real picture of the purchasing firm's track record.

CREDIT ISSUE : Now in any given state of the system, once the regular purchasers demands are met, the remaining values suggested purchase value of the control input, can either be ignored or used to purchase yarn and issue to some bad debtors as well according to the following method

Share Credit For Bad Debtor X = Sum of all bad debtor score/ Company X's score

Since the lower the score the better, the company with the lowest score will get the highest share.

Share in remaining yarn purchased and weaved into the final product poplin = Share Credit/Sum of all share credits. With the previously used table as an example, the Share credit formula and proportion of poplin cloth remaining given in credit to the bad debtor is given below:

Sum of all bad debtor credit score = 1.14 + 1.21 + 1.24 + 1.33 = 4.92. Therefore,

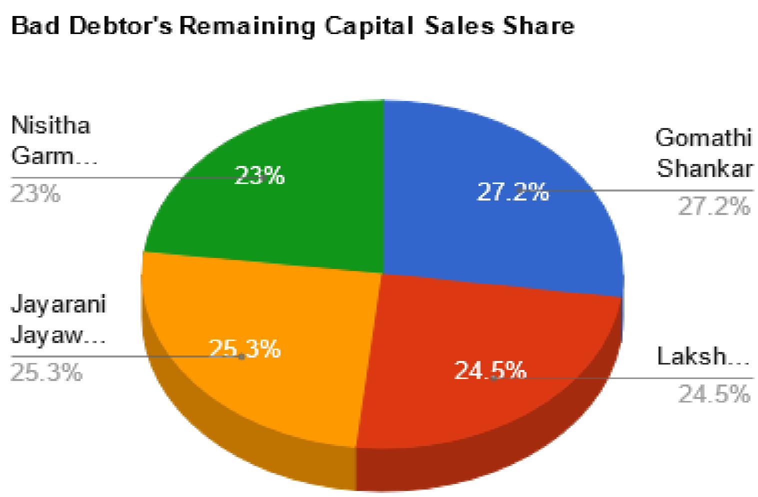
DEBTOR	SHARE CREDIT
Gomathi Shankar	4.92/1.14 = 4.35
Jayarani Jayawani	4.92/1.21 = 4.06
Lakshmipriya Textiles	4.92/1.24 = 3.96
Nisitha Garments	4.92/1.33 = 3.69

Sum of Shared Credit = 4.35 + 4.06 + 3.96 + 3.69 = 16.06

DEBTOR	SHARE OF SALES
Gomathi Shankar	4.35/16.06 = 0.272
Jayarani Jayawani	4.06/16.06 = 0.253
Lakshmipriya Textiles	3.96/16.06 = 0.245
Nisitha Garments	3.69/16.06 = 0.23

So, from the above table, we can infer that after regular non-defaulters are satisfied with the state space system's suggested control input, the bad debtors may get the above share of sales in cloth or the business person may hold it in reserve.

When giving a share of sales to a bad debtor, the business owner could strictly follow the above share rule to cut losses. So 'Gomathi Shankar' will get a share of 27% of the remaining capital's worth of poplin and so on. A pie chart is attached below for visualization.



Bad Debtor's Remaining Capital Sales Share

It is once again important to note that while this is a postmortem analysis, a live credit score update system must work in tandem with the state space model in order to streamline the decision making process. This too involves keeping frequently updated records to get the most accurate credit score

And hence our business analysis process is complete to meet all our objectives.

Access to a rough version of the analysis conducted in a Colab file has been attached here :
Google Colab File Link : State Space Analysis - BDM
Note : The list of references for these methods are attached towards the end of this document.
The necessary data is available to modify and experiment in the Colab file from this link :
Data Folder

INTERPRETATION AND RECOMMENDATION:

1. **RECOMMENDATION :**
   The business faces the issue of analyzing an extremely complicated and erratic dataset. Simple regression analysis and curve fitting would provide only haphazard curves that do not reflect the nature of the curves of the real data. From this arises a need to automate the decision making process for the

This may be rectified by implementing state space analysis methods that can compensate for a whole set of unidentified factors by modeling the business as a control system, with feedforward matrices which can function as feedback matrices upon inversion of the sign of parameters to establish a system equation upon which one might perform predictive analytics.

To efficiently utilize this system, the business person will only need to have the knowledge of his state variables, which are the different factors that influence the state of the system in his business.

By simply plugging in the state vector data, he would be able to accurately predict the expected turnover in the long run, the necessary yarn purchase he would need to make in advance and the expected error range or fluctuation in his expected results and state variables in the future in that financial year.

STEPS TO FOLLOW:

To utilize this system to its fullest extent, the following steps/measures must be followed or taken into account by the decision maker at any given time.

1. Maintain an always up-to-date record of the state variables. This includes all of the expected sales/accounts receivable at the end of the individual day.
2. The state space model was trained for each day's regression valued output and to get the most out of the model. It would require meticulous day-to-day record keeping
3. Utilize the state variables in the first state equation to choose a conservative input.
4. Since the state vector derivative for the long term in a fiscal year is already determined with the model that has been built.
5. Now use the second equation and substitute the control input derived from Step 3 and the knowledge of state vectors at the time of the decision making process and predict the output, expected sales
6. Since state space models are sensitive to changes in the context of the state vector derivative, it is recommended that a the model be retrained and updated every six months

7. This is due to the fact that there may be new customer or there may have been new events that drastically affect the state variables
8. These events could be a supply chain issue that alters the yarn price, which would mean the regression model would have to be updated drastically. This would need to change the state space model as well
9. Or it may be issues such as a catastrophic event in the business or even disruptions to long term high demand customer businesses due to competitors
10. Subsequent to the periodic training and analysis, it is vital for the business owner to keep in mind that state variables and the current model is restricted by the data available.
11. It is meant to give a macroscopic view of the system while serving as an indicator of whether the minor business decisions would have a major impact on the business or is inconsequential in the long term as the model is usually resilient to such minor changes.
12. This is clearly observable with the high state derivative vector, which has a high value and is around 10% of the value of the business.
13. So it is safe to assume that there wouldn't be major changes since the business operates in far higher profit margins
14. Hence, yarn price forecasting, cash-flow optimization could be reasonably managed and incorporated within this system with these models for efficient decision making.
15. Now that the loss function has been integrated into the state space system, incorporating the bad debt recovery into the pipeline of operations and judging it's effect on future operations now becomes an easy task.
16. Staggered credit is entirely dependent upon the

2. POINTS TO NOTE:

1. It is vital to remember that the chosen state variables with available data will never be sufficient to model the real world system's state space in its entirety.
2. Factors such as competitor influence, debtor's internal variables, capacity utilization of the firm etc., have not been taken into account.

3. A more perfect system will have a higher number of state variables, including but not limited to the existence of variable costs, time delays records of dying and weaving (which are outsourced by this firm), etc.
4. Since a state space system gives a broad understanding but is sensitive, careful deliberation must be made and the right call at the right time to retrain the model with new data if drastic changes occur in the business scene and not always stick to the bi-annual review of the control system equation and its performance.
5. The model development has relied on regression and descriptive analysis of the data and the business person's decades of experience while building the model. Newer state variables might newer estimator methods before being formulates and trained on the state space equation

3. FUTURE SCOPE OF IMPROVEMENT:

1. Different state variables with more data and better knowledge of internal workings of the customers and outsourcing companies could be modeled with different algorithms.
2. Black-Scholes PDE could be employed to deploy a Merton Credit Risk Model credit rating system for the all businesses that purchase poplin from this poplin
3. Wavelet Transforms with appropriate wavelet function and Kalman filters could be employed to denoise the regression model before feeding into the state space system for training and decision making.

LIST OF REFERENCES:

1) State Space Analysis State-Space Approaches for Modelling and Control in Financial Engineering: Systems Theory and Machine Learning Methods - by Gerasimos Rigatos

2) Control Systems Engineering - by Norman Nise

3) Optimal Control Theory, Applications to Management Science and Economics - by Suresh P. Sethi

4) Systems Science: Theory, Analysis, Modeling, and Design - by George E. Mobus

5) Near Optimal Control in Ride Hailing Platforms with Strategic Servers - Sushil Mahavir Vera et al. (Reference paper for a similar method previously used in a different business)