Data Mining Project Proposal

Noelle Hartman
Department of Data Science
Florida Polytechnic University
Lakeland, Florida, USA
nhartman2951@floridapoly.edu

Anastasiya Dmytryk
Department of Computer Science
Florida Polytechnic University
Lakeland, Florida, USA
admytryk9341@floridapoly.edu

I. Motivation

For centuries, fashion and apparel has been a booming industry, and over this period, changes in science and technology have revolutionized it. One of the most prominent changes to the fashion industry is undoubtedly the advent of mass production capabilities unfolding in the Industrial Revolution, which has contributed to the now huge global market size amounting to nearly two trillion dollars as of 2025 [6]. Obviously, with an industry this large, problems relating to sustainability, design, optimization, and forecasting are bound to be plentiful. Thankfully, with the advent of machine learning, many of these problems can be investigated and potentially mitigated. According to a review published by the 2024 ASU International Conference in Technologies for Sustainability and Intelligent Systems, machine learning can be applied to the following sub-fields in fashion: design, production, consumer experience, and sustainability [6]. With such a wide range of applications, it is safe to assume that fashion technology, especially pertaining to machine learning, is an emerging and complex field in need of further study.

Of the subfields mentioned above, the issue of production stood out as especially pertinent. This is because without production, the other subfields (design, consumer experience, and even sustainability) fade into obsolescence, as their importance is dependent on the existence of a tangible product. In fact, production is such a critical matter of interest that last year Dmytryk investigated optimizing the cutting processes of fabrics in textile manufacturing using machine learning methods and found that the model was successful at the optimization task and was effective in reducing fabric waste. Production of textiles is obviously dependent on the availability of material commodities, and one of the most popular commodities used in clothing, cotton, was particularly interesting.

Cotton is an ancient fiber and has extensive use cases. Not only can this fiber be used to create comfortable clothing, but its seeds can also produce edible oil, and these seeds are also used in livestock feed [7]. Although cotton cultivation is frequently criticized as being unsustainable, there has been plenty of ongoing research into improving the sustainability of cotton cultivation. For example, according to a review from the Department of Biological and Agricultural Engineering at Texas A&M University, more sustainable sowing, harvesting, and irrigation techniques are being studied and are showing varying degrees of promise [7]. Furthermore, cotton has also shown extensive applications in medicine. Some of these applications include surgical clothing, wound dressing, suture materials, and even drug delivery [10]. These applications are possible due to the high cellulose fiber content, large surface area, and high gas permeability of cotton-based materials [10]. Clearly, cotton is a staple commodity and is the subject of progressive and ongoing research, indicating that this material will continue to be an integral component in the manufacturing process of textiles for many years to come. Also, cotton has plenty of uses beyond simply textile production in the fashion industry, so any insights into this commodity, especially regarding its pricing, have the potential to be generalized across fields and could cause a wide range of impacts.

Speaking of cotton pricing, commodity prices, especially commodity price volatility (CPV), have always been a puzzling issue facing companies, from oil and gas to meat products [3]. Many studies have been published regarding this problem, and they mainly focus on forecasting commodity prices accurately and developing supply chain strategies to minimize the impact of price volatility on a company's profits. According to a paper published in the IEEE Transaction on Engineering Management, "It is clear that any volatility in the price of raw materials (inputs) leads to fluctuation in end products' (outputs) price" [3]. The severity of fluctuation in the outputs' price is also directly related to the severity of fluctuation in the inputs' price according to the researchers. It is also noted that developing nations face the greatest challenges when it comes to CPV [3]. To help companies better cope with this oftentimes stressful and confusing situation, the researchers developed a machine learning model to better understand which raw materials most dramatically affected the cost of goods sold (COGS) value, or the actual price of the output for consumers. The company of interest in this study produced machinery, meaning that there were multiple manufacturing steps, each requiring their own share of raw materials. Using the model, the researchers conducted a sensitivity analysis to see which raw materials at which manufacturing stage caused the greatest upset in the COGS value, and found that "once the price of one particular raw material changes to some extent [in the second stage of manufacturing], then the price of other associated raw materials in that level is also perturbed to some degree" which manifested in changes to the COGS value accordingly [3]. This led the researchers to advise that the company "needs to be vigilant about such a volatility in order to become robust against it" [3], and to become robust, the company obviously must implement strategies to mitigate this COGS fluctuation. Thankfully, researchers have addressed this strategy concern. According to an article from Supply Management: An International Journal, researchers propose two strategies to better reduce the negative impacts of CPV. These strategies are Total Cost of Ownership (TCO), and Real Options Valuation (ROV) [8]. In a nutshell, TCO pinpoints where supply chain interventions (e.g. changing the material supplier or investigating an alternative material) will have the most cost impact, and ROV helps decision-makers compare strategies and understand how external changes (like market shifts) might affect performance overall [8]. Combined together, these strategies are crucial in the company's sensitivity analysis of the COGS value [8]. With these two papers in mind, it is clear that the price of raw materials has a significant impact on the price of the manufactured product, and manufacturing companies invest much time and study into the impacts of CPV; in other words, this is a non-trivial issue and is worthy of further investigation.

Given that CPV is such a pervasive issue, it is safe to assume that this problem must also be prevalent in the fashion industry, particularly in textile manufacturing. We choose to focus on the fashion industry due to its growing technological sector and the fact that one of the contributing students to this project has prior experience in the field. In order to better understand and predict the price fluctuations of textiles, it is natural to pinpoint one of the main commodities used in their manufacturing process, cotton. The relevance of cotton as a material is also justified by the papers mentioned prior, which detail the ongoing research investment into its cultivation as well as its plethora of important applications beyond simply textile manufacturing. Our goal is to determine a correlation between global raw cotton prices and textile consumer prices. After and if such a correlation is established, we will investigate what factors in the cotton cultivation process are most likely to contribute to the CPV to better forecast the raw material cotton price. With this study, we can contribute to the emerging field of fashion technology, and our focus on cotton CPV can also assist other fields where cotton is an important material.

II. Related Work

Fashion technology is a growing field, and machine learning can be applied to many facets of the fashion industry, such as design, production, consumer experience, and sustainability [6]. As mentioned in the motivation, the main interest of this project will be the production aspect, where we will be establishing a correlation between global raw cotton prices and the price of textiles. If such a correlation exists, we will investigate what factors in the cultivation process hold the most sway over the CPV of cotton. Thankfully, we will not be alone in this endeavor, as there is already a sizable body of related literature.

The first article in this body is written by a researcher hailing from Bina Nusantara University, and the topic of interest is how increases in the prices of raw materials influenced the production output of textile manufacturers in Indonesia during the Covid-19 pandemic [1]. It is noted that during this time, the price of raw materials for these manufacturers had increased by 30%. The researcher's null hypothesis stated that there would be no correlation between the raw material price increase and the production output, and the alternative hypothesis stated that there would be a correlation. The researcher deployed an online survey questionnaire to 100 textile manufacturers, and from the results determined via simple regression that there was a correlation between the price of raw materials and output. The nature of this correlation was slightly paradoxical, as it conveyed that raw material price and output were directly related. This led the researcher to the optimistic conclusion that the textile business was very likely to survive the Covid-19 pandemic. Although this article offers interesting insights about the price of raw materials and the production output of textiles, it does not offer much color on how the price of raw materials influenced the price of the textiles. This is a gap that our project will attempt to fill.

Another piece of relevant literature shows how data mining techniques and machine learning can be used to predict sugarcane yield [5]. In this study, researchers gathered data about relevant cultivation factors such as precipitation, temperature, and the composition of the planting soil from a sugarcane mill. They noted that any features that were obtained near the harvest period were discarded because these features probably would not have had a significant impact on the growth of the crop. With this data at hand, the researchers employed 6 different modeling techniques such as feature engineering, feature selection, and hyperparameter tuning, to produce 66 modeling combinations for testing. After testing, it was found that the average model error was about 6.42 megagrams per hectare (Mg/ha), with the best having an error of 4.11 Mg/ha and the worst having an error of 9.00 Mg/ha. The baseline model simply predicted the expected yield based on the number of prior harvests, and this model had an error of 9.86 Mg/ha, so the machine learning models, even at their worst, far outperformed the baseline. There were also some interesting findings regarding the tweaking of the modeling techniques. For instance, tuning the models reduced the error by 1.17 Mg/ha and feature engineering reduced the error by 0.64 Mg/ha. Surprisingly, feature selection (in this case removing 40% of the features) actually made the error increase by 0.19 Mg/ha. This paper, although not focusing on cotton yield specifically, provides valuable guidance on how to tune a model in the context of crop yield problems. The fact that feature selection increased the error implies that most of the data features the researchers gathered were relevant in predicting the crop yield. For our particular project, after we determine the cotton price and textile price correlation, we plan on attempting to predict cotton prices, which will likely hinge on cotton yield forecasts. It is likely that we will be developing these forecasts using available cotton cultivation data, and our feature selection for this predictive model can be guided by the feature selections made in this paper.

The next piece of related material is from the Helsinki School of Economics. The paper mainly grapples with the problem of inventory valuation and costing, two very important metrics for a company's finances [2]. The author stresses that volatile prices on raw materials are a major problem for making accurate estimates about these important financial metrics, and to help solve this problem, the author ran a simulation model to test different raw material inventory valuation methods. For completeness, the author also ran a scenario analysis to check how effective each raw material valuation method was in determining an accurate inventory and cost valuation. With these raw material valuation methods, the company used for the simulation saw product costing errors reduced by 14%, and inventory value fluctuations and inconsistencies were reduced by 7%. Although this paper does not directly tie into our textile pricing problem, it highlights how disruptive volatile raw material prices can be on a company, and this level of disruption can lead to loss of revenue, raising the consumer price of the company's product. This paper was also written in 2009, revealing a lack of recent developments in the field of commodity pricing. We hope to contribute to filling this gap by providing more modern solutions to a similar issue.

An additional and pertinent piece of literature is a paper published in the journal Neural Computing and Applications, and it explores how machine learning models can predict the prices of agricultural commodities. The paper uses the key variables of supply, demand, yield, and year in the model. According to the authors, previous related studies have not adequately accounted for supply and demand data for a specific crop even though these features are crucial for price predictions. The authors also show how machine learning models outperform other methods for predicting the price such as time series and statistical regression, and that deep analysis of training data improves the forecasting accuracy. The authors stress that in order to accurately predict the consumer price of a crop, market factors, such as supply and demand, must be considered. They even suggest adding more socioeconomic features to their model, such as disasters, inflation, and the availability of alternative crops, in order to obtain more well-rounded predictions thoroughly accounting for all potential factors. This study will be very influential in our project, specifically in the predication of cotton prices, because it shows that external socioeconomic and market factors (e.g. supply, demand) are just as important as the production factors (e.g. yield, labor costs). In other words, the paper has provided us with expert guidance on which factors to consider when determining the price of a crop, and how to work with the data related to these factors.

The final paper in the body of related literature is very specific to our topic and offers very useful data and insights about cotton yield in Karnataka, India. The main objective of the study was to show how artificial neural networks (ANN) performed better than statistical models. The network was trained using a long-term dataset spanning 1990 to 2023, with prediction errors generally within ±10% during key growth stages. The authors also used evaluation metrics on the model such as RMSE, nRMSE, and EF, and these all fell within acceptable ranges across most of the districts studied. The authors also emphasized that weather fluctuations and temperature are some of the key factors in determining cotton yield, and the ANN was able to determine which of these weather factors had the greatest influence on cotton yield. A seemingly trivial weather factor, morning relative humidity and its interaction with maximum and minimum temperatures, had a major effect on cotton yield in most districts. It was also found that the impact of certain weather factors was dependent on the district, specifically the local soil and crop management practices. This work is especially valuable for our project because it highlights the usefulness of machine learning in modeling crop performance under rainfed conditions, which closely resemble the environments where much of the world's cotton is grown. It also reveals the complexity of predicting cotton yield in India, as each district responds to different weather factors a bit differently. Since our project also involves forecasting raw cotton prices, understanding how weather-driven yield variations affect supply could improve our cotton price prediction model. The feature importance analysis in this study can also guide us in selecting relevant environmental variables when attempting to link yield and price volatility.

As for data sets, so far we have collected three resources: Producer Price Index by Commodity: Textile Products and Apparel: Spun Cotton Yarns [12], Producer Price Index by Commodity: Textile Products and Apparel: Finished Cotton Broadwoven Fabrics [13], and Global price of Cotton [14]. All of the datasets were sourced from the U.S. Bureau of Labor Statistics.

III. Proposed Approaches

This study will focus on figuring out if there is a correlation between raw cotton price volatility and the final textile price globally. Furthermore, the study will investigate the possibility of forecasting textile prices using the raw cotton price and does raw cotton Granger-cause textile prices [11].

After an intensive search of many databases, the team was unable to find any data sets that would directly show case the price of raw materials and the price of textiles. However, some data was found that can be transformed to a standardized form for the prediction. The data sets used will be as follows:

- **The global price of cotton**: https://fred.stlouisfed.org/series/PCOTTINDUSDM
- **Producer Price Index (PPI) – Finished Cotton Broadwoven Fabrics**: https://fred.stlouisfed.org/series/WPU034201
- **Producer Price Index (PPI) – Spun Cotton Yarns**: https://fred.stlouisfed.org/series/WPU032601

These data sets are all time series data; however, they all have slightly different ranges. For this study the year range will be January 1991- July 2025.

As there are two data sets, the Finished cotton and the Spun Cotton, we will conduct two experiments to determine the correlation of the global price to both separately. And then test what impact raw cotton and spun cotton have of finished textile. The variable of this experiment will be described as follows:

- **Raw Cotton vs Finished Cotton Fabric**
  - X_t: Raw cotton price at month t
  - F_t: Finished textile index at month t
- **Raw Cotton vs Cotton Yarn**
  - X_t: Raw cotton price at month t
  - Y_t: Cotton yarn index at month t
- **Finished Cotton vs Raw Cotton and Cotton Yarn**
  - X_t: Raw cotton price at month t
  - F_t: Finished textile index at month t
  - Y_t: Cotton yarn index at month t

In the data sets the cotton price is measured in cents per pound and the textile price and yarn price is measured as an Index 1982=100. Because of that all the data will be represented as Index at year y = 100. The data avaliable is also spanned across different dates, for this experiment the window was selected from Jan 1991- July 2025.

The year for the index will be average price in 1991. All the data will be transformed to follow this form.

To begin each data set will be plotted individually to determine the shape of the data, along with the pairwise scatter plots to see how the data looks together.

If the data has a linear relationship the person correlation coefficient will be calculated for all pairs.

The data will be split into 80-20 with 80 being the amount of training data, then the data will be split using Time Series Split cross validation to ensure there is no date rearrangement.

IV. Planned Experiments

The data used is a time series data, and to analyze the relationship different regression models will be used.

To test the one-to-one influence of raw cotton to yarn, raw cotton to textile, and yarn to textile Simple Linear Regression will be used.

- **Raw Cotton to Textile:**

$$Y_t = \beta_0 + \beta_1 X_1 + \epsilon_t$$

- **Raw Cotton to Yarn:**

$$F_t = \beta_0 + \beta_1 X_1 + \epsilon_t$$

- **Yarn to Textile:**

$$F_t = \beta_0 + \beta_1 Y_1 + \epsilon_t$$

- **Multilinear Equation:**

$$F_t = \beta_0 + \beta_1 X_t + \beta_2 Y_t + \epsilon_t$$

- F_t — finished cotton textile index at time t
- X_t — raw cotton index at time t
- ϵ_t — error at time t
- Y_t — yarn index at time t

Granger causality to see if raw cotton time series can help predict finished textile time series.

To determine the lag size a [1, 12, 24, 26, 48] lag will be used, and their BIC (Bayesian Information Criterion) scores will be compared to determine the best fit.

The restricted model for textiles will follow this equation:

$$F_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i F_{t-1} + \epsilon_t$$

The model for the full prediction will be:

$$F_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i F_{t-i} + \sum_{j=1}^{p} \beta_j X_{t-j} + \sum_{k=1}^{p} \theta_k Y_{t-k} + \epsilon_t$$

With this equation the prediction is expected to cover is there a relationship between raw cotton and finished textile, can we predict the price of the textile based on cotton, and what variables are actually a driving change in textile price.

To predict the relationships a Garger Causality test and a regression coefficients significance test will be performed to determine the correlations

To test how well the future values are predicted a Mean absolute Error and Mean Squared Error and Root Mean Squared Error tests will be performed.

References

[1] R. R. Dwitomo, "Relationship between the price of raw materials of fabric with the amount of production produced," *PSSJ*, vol. 1, no. 1, pp. 12–15, Jul. 2021.

[2] H. Suviolahti, "The influence of volatile raw material prices on inventory valuation and product costing," Helsinki School of Economics, Jun. 2009.

[3] H. Moheb-Alizadeh and R. Handfield, "The impact of raw materials price volatility on cost of goods sold (COGS) for product manufacturing," *IEEE Transactions on Engineering Management*, vol. 65, no. 3, pp. 460–473, Aug. 2018, doi: 10.1109/TEM.2018.2796447.

[4] M. N. Thimmegowda, M. H. Manjunatha, H. Lingaraj, et al., “Comparative analysis of machine learning and statistical models for cotton yield prediction in major growing districts of Karnataka, India,” *Journal of Cotton Research*, vol. 8, no. 6, 2025. [Online]. Available: https://doi.org/10.1186/s42397-024-00208-8

[5] F. F. Bocca and L. H. A. Rodrigues, “The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling,” *Computers and Electronics in Agriculture*, vol. 128, pp. 67–76, 2016. [Online]. Available: https://doi.org/10.1016/j.compag.2016.08.015

[6] K. Dhiwar, “Artificial intelligence and machine learning in fashion: Reshaping design, production, consumer experience and sustainability,” in *Proc. 2024 ASU Int. Conf. Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, Manama, Bahrain, 2024, pp. 1766–1775, doi: 10.1109/ICETSIS61505.2024.10459436.

[7] A. Adeleke, “Technological advancements in cotton agronomy: A review and prospects,” MDPI AG, Feb. 23, 2024. doi: 10.20944/preprints202402.1342.v1.

[8] B. Gaudenzi, G. A. Zsidisin, and R. Pellegrino, “Measuring the financial effects of mitigating commodity price volatility in supply chains,” *Supply Chain Management: An International Journal*, vol. 26, no. 1, pp. 17–31, Jan. 2021.

[9] M. K. Mohanty, P. K. G. Thakurta, and S. Kar, “Agricultural commodity price prediction model: A machine learning framework,” *Neural Computing and Applications*, vol. 35, pp. 15109–15128, 2023. [Online]. Available: https://doi.org/10.1007/s00521-023-08528-7

[10] M. Shahriari Khalaji, A. Alassod, and Z. Nozhat, “Cotton-based health care textile: A mini review,” *Polymer Bulletin*, vol. 79, pp. 1–24, Jan. 2022, doi: 10.1007/s00289-021-04015-y.

[11] H. Lin, M. Ren, P. Barucca, and T. Aste, “Granger causality detection with Kolmogorov-Arnold networks,” *arXiv preprint arXiv:2412.15373*, 2024. [Online]. Available: https://arxiv.org/abs/2412.15373

[12] U.S. Bureau of Labor Statistics, “Producer Price Index by Commodity: Textile Products and Apparel: Spun Cotton Yarns [WPU032601],” FRED, Federal Reserve Bank of St. Louis, Jul.2025. [Online]. Available: https://fred.stlouisfed.org/series/WPU032601

[13] U.S. Bureau of Labor Statistics, “Producer Price Index by Commodity: Textile Products and Apparel: Finished Cotton Broadwoven Fabrics [WPU034201],” FRED, Federal Reserve Bank of St.Louis, Jul.2025. [Online]. Available: https://fred.stlouisfed.org/series/WPU034201

[14] International Monetary Fund, “Global price of Cotton [PCOTTINDUSDM],” FRED, Federal Reserve Bank of St.Louis, Jul.2025. [Online]. Available: https://fred.stlouisfed.org/series/PCOTTINDUSDM