2019

Data Analytics for Supply Chain Management

MACHINE LEARNING APPLICATIONS IN E-COMMERCE, DELIVERIES & PRODUCTION

ATIT BASHYAL, TANASORN CHINDASOOK, JANDRA FISCHER, HAMZA INTISAR, MARK KOERNER, PETER-SLEIMAN MANSOUR

JACOBS UNIVERSITY BREMEN

28. NOVEMBER 2019

Contents

List of Figures

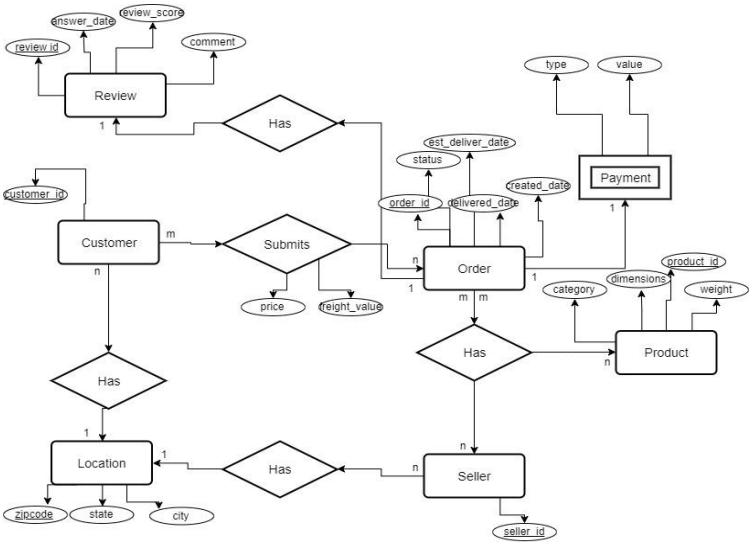
1. Introduction

In the context of supply chain and in order to get familiar with supply chain related datasets, we were asked to develop use case scenarios based on certain open dataset. The first step of this project was to propose datasets that were publicly available, easy to understand, interesting to work on, challenging to analyze and concern real world scenarios. Therefore, and in order to be able to cover more than one topic of the supply chain, different analysis methods and various scenarios three datasets were chosen from the online platform Kaggle. While the first dataset covers the E-commerce business and shells data regarding sales. The second dataset lists details about cashew truck deliveries and focuses on the procurement part of the supply chain. Finally, the third dataset covers the production topic and catalogues real world data obtained from a floatation plant. This report is divided into three sections, in which we will take an extended look at each dataset, put it into the right context and attempt to provide solutions.

2. E-Commerce Data Analytics: OList Brazil

Olist is a Brazilian e-commerce platform founded in 2015 that sells a wide variety of products from different shops on the main online marketplaces in Brazil. The dataset comes from the Kaggle website and concerns the sales part of Olist's Supply Chain. It lists data of more than 100,000 orders made in the years 2016 until 2018 and contains a total of 9 tables. The table descriptions for the Olist dataset can be found in Appendix 1.

In total, these 9 tables contain 51 variables. However, some variables are duplicates in different tables. For instance, all three variables Zip Code, City and State appear in the three tables Customers, Sellers and Geolocation. This poses a data integrity issue in storage, as the values in all three tables would have to be updated if a seller or customer changed locations. A suggested improvement is that each location be stored by a unique identifier (location ID) and the duplicate locations in the other tables should be referred to by their location ID, thus creating a foreign key reference in both tables instead of duplicate data issues. The general ER diagram for the Olist dataset is shown in Figure 2.1.



1 Figure 2.1 ER diagram for the Olist dataset

Finally, we suggested adding three attributes related to the date the supplier issued the shipment, the date the truck actually delivered the shipment and the classification of the supplier.

By looking at the outputs of our models, we can assume that none of them predicts the output accurately. Nevertheless, by creating strategic alliances with our top suppliers, we can expect a constant data flow and maybe develop new features to measure. Finally, with a better structured dataset and more information about the shipments, the model created will be able to better predict the expected outcome.

4. Product Quality Control: Iron Ore Production

This particular dataset contains manufacturing process data from a real world iron mining floatation plant. It contains 24 columns describing different aspects of the flotation process in iron ore mining. This process is a standard procedure to further concentrate the iron ore. The attribute descriptions can be found in Appendix 5.

4.1. Suggested Dataset Improvements

The current dataset contains some iron ore production values in hour intervals and some in 20 second intervals within the same table. The hourly values are simply repeated 180 times, which is also true for the hourly timestamps. Since there are no precise timestamps or disclaimers for the values with a frequency of 20 seconds, it was up to us to figure out which ones were in which frequency and to assume that the intervals were in the correct order for every hour. In order to avoid these assumptions, it would be good if similar datasets in the future provided precise timestamps and appropriately named variables for each frequency or just split the data with different frequencies into separate tables with the ability to join them on the time indexes.

4.2. Supply Chain Context

The Iron Mining Process dataset is limited in its applicability to different supply chain scenarios. The stated main goal on the Kaggle website is production control, which is the only scenario it is properly suited for. Production planning is the only other slightly relevant context but since the dataset contains percentage contents of the final product instead of numeric quantities, it does not allow for this application at the most basic level. Thus, it is only suited for production control, and specifically only for quality control and monitoring.

In manufacturing defect prediction, the main variables usually include values measuring the quality of the input materials as well as other relevant measurements throughout the process[13]. In case of the Iron Ore Mining dataset, the most important variables would thus be the starting purity measures such as % Iron and Silica Feed as well as other direct process measurements such as the Flow and Ore Pulp variables. Indirect process measurements such as the air flow and froth level could also have an impact, which will ultimately be determined throughout the modelling process.

[13] (Santos, et al., n.d.)

4.3. Scenario Development

Currently, the engineers at the plant do not have a convenient and reliable way to measure the iron ore impurity, i.e. the quality of their product. If an engineer wants to assess the quality and contents of the iron ore at the end of the floatation process, the contents of the ore have to be measured in a lab which takes about an hour. This means that engineers can only take actions to ensure proper product quality with at least an hour delay and only in case a sample was even chosen for testing to begin with. Thus, the plant's engineers lack a proper way to diagnose and continuously monitor product quality, which could lead to poor product quality and even cases where the product cannot be sold for its intended purpose.

The goal for analyzing this dataset is hence to provide the engineers a data-driven solution for monitoring the product quality during the iron ore concentrate production process. A successful impurity model would allow the engineers to respond to potential cases of poor product quality in a more timely and organized manner, ultimately helping the plant by improving product quality on average and preventing long periods of poor production quality.

We attempted to model the impurity with two different approaches. Initially, we tried different regression methods in order to directly forecast the percentage of Silica Concentrate as part of the final Iron Ore Concentrate. Moreover, after looking at the distribution of the Silica Concentrate values, we decided to try classification methods in order to identify "impure" or "pure" batches. The classification output could then be used to trigger a warning to engineers instead of showing the potentially imprecise or misleading regression forecast.

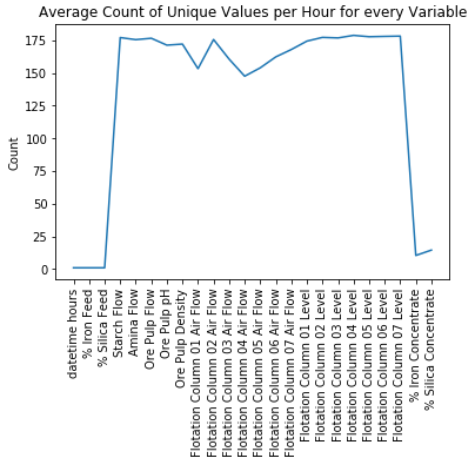
4.4. Data Exploration and Preprocessing

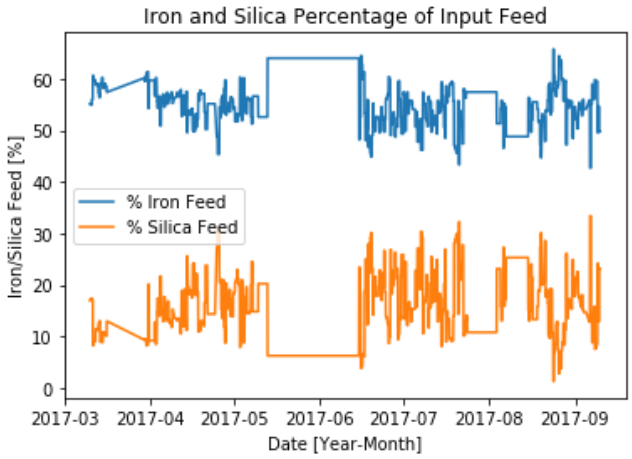
The goals for this step are to gain a deeper understanding of the dataset at hand and prepare it for the modelling step. Upon initial inspection, the dataset contains no explicit missing values and a little more than 737.000 rows. Since the dataset consists of only one table, we did not need to transform its structure initially.

As previously mentioned, some of the variables are provided in hourly frequency and some in 20 second frequency, so the first step was to determine said measurement frequency. Figure 1 shows how many unique values per hour each variable averages. It seems that only % Iron Feed, % Silica Feed, % Iron Concentrate and % Silica Concentrate are provided in hourly frequency, as the rest of the process measurements only contain very few repeating measurements on average. It is important to note, however, that both Concentrate variables have an average unique count of above 1, which indicates some inconsistencies in the data.

As the majority of the variables are not in hourly frequency, we decided to create a column detailing the exact measurement moment, assuming the observations within each hour were in the correct order. Before proceeding, we also tested which hours had less than 180 records, which is the amount of 20 second intervals in an hour. Throughout this process, we noticed that two hours contained less than 180 records and that data for some hours was missing from the dataset entirely. Additionally, some hours for % Silica Concentrate, our intended target variable, contained exactly 180 unique values, seemingly due to an interpolation procedure between the value of the previous and upcoming hour. We decided to remove those hours from the forecasting dataset. This exploration can be seen in Figure 4.1.

When individually graphing all variables as a lineplot, both Iron and Silica Feed stood out for similar reasons. Both graphs show multiple plateaus, where the values are constant for an extended period of time as shown in Figure 4.2. After investigating further, we decided to remove the corresponding rows from the dataset as well, as the input components of the iron ore seem like important factors in this case. In other cases, we could have also considered either excluding the variables completely or keeping them as is, but the perceived value of those two variables in this context shaped our decision.

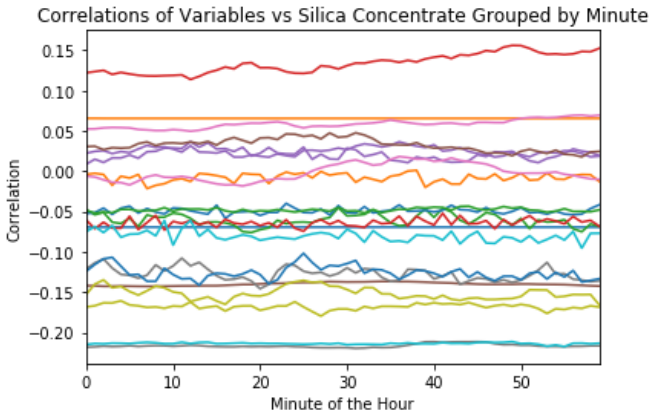




18 Figure 4.1: Lineplot of average unique values per hours & Figure 4.2: Time Series Plot of % Iron Feed and % Silica Feed for the entire dataset.

To further delve into the relationships between the variables, we wanted to examine the correlation between the features in order to inform our modelling decisions. We found that there were no helpful, significant correlations. Appendix 6 shows a pairplot of select variables, which includes scatterplots between each variable and a histogram to show the distribution of each variable along the diagonal. The only apparent patterns indicate a relationship between % Iron Feed and Silica Feed as well as Iron Concentrate and Silica Concentrate, which are to be expected as both are percentage contents of the same material.

Our last hypothesis was that - assuming the data points were ordered correctly and measurements were usually carried out around the same time - each variable should exhibit a higher correlation with the % Silica/Iron Concentrate around the time the measurements were usually taken. Thus, we decided to group the dataset for minutes within the hour, and examine the correlation within those subgroups. The correlations for all variables are fairly low and do not follow a significant hourly pattern overall, as shown in Figure 4.3. As a result, the measurements seem to be either taken at random throughout the hour or the moment of measurement has no impact on the correlation with the process variables.



19 Figure 4.3: Lineplots depicting correlation between all individual variables and % Silica Concentrate grouped by minutes of the hour

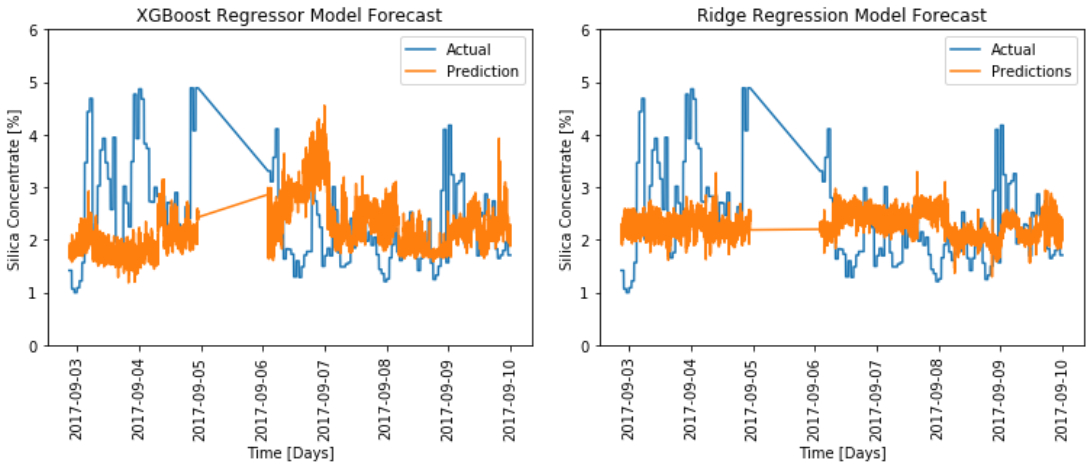
4.5. Data Analysis

After data exploration and cleaning, the next step was to attempt to model the Silica Concentrate based on the input and process variables. Overall, the dataset seemed fairly uncorrelated, meaning it could prove to be difficult to produce accurate models. In addition, the inherent nature of the values and their measurement frequencies presented a cause of concern, as we had to decide which frequency to use for forecasting. We ultimately decided to stick with the lower 20 second frequency, as this allows us to use all of the remaining data. Initially, our goal was to numerically predict the % Silica Concentration using tree-based and regular regression algorithms. Tree-based algorithms seemed especially promising in this case considering the low linear correlation between the variables.

We started the modelling attempts using the XGBoost tree-based Regressor, which uses gradient boosting in order to find the optimal tree structures[14]. After a little bit of experimental parameter tuning, we decided to fit the model on the first 130 days worth of data and predict the rest, as shown in Figure 4.4. While the predictions seem generally close, they do not follow the actual patterns and do a poor job of correctly predicting the spikes in % Silica Concentrate. Comparing the accuracy measures, a value for RMSE of 1.14 and MAE of 0.87 for values ranging between 1 and 5 is very high. The RMSE is the square root of the average squared error, the MAE is the mean absolute error value.

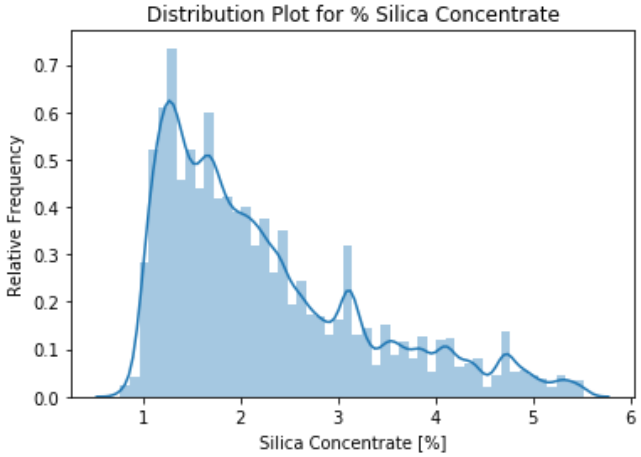
Considering such disappointing results, we decided to fit a Ridge Regression model for comparison, Ridge Regression is similar to Linear Regression, except that it includes a regularization term in its loss function to prevent overfitting[15]. As compared to the XGBoost modelling attempt, the Ridge Regression shows an even lower capacity to correctly forecast the important outliers as shown in Figure 4.4. Even though the accuracy measures for Ridge Regression are slightly better than for XGBoost with an RMSE of 0.97 and a MAE of 0.75, they are hardly inspiring.

[14] (Chen, n.d.)
[15] (Hoerl & Kennard, 2000)



20 Figure 4.4: Time Series Plots depicting the actual values for % Silica Concentrate and the predicted values from the XGBoost Regressor model Ridge Regression model respectively

As numeric prediction did not yield the necessary results, we decided to try a different approach. After looking at the distribution of the % Silica Concentrate target variable shown in Figure 4.5, we noticed a left-skewed distribution with a long tail on the right side, indicating a fairly substantial amount of high impurity cases as the percentage of Silica in the Concentrate increases. As predicting the higher values in Silica percentage is of utmost importance, we chose to implement a classifier which would label a sample as impure if it contained more than 3% Silica based on the distribution plot. Our goal was not only to produce an accurate classifier, but most importantly to produce a classifier that was accurate in predicting impurity.

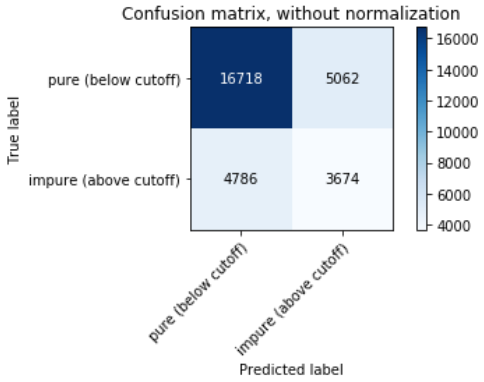


21 Figure 4.5: Histogram and Distribution Plot of the % Silica Concentrate Variable

Again, we decided to try one tree-based and one standard regression method. Due to its inherent nature as defect detection, the classes were fairly unbalanced. Initially, there were close to 4 times as many pure observations as there were impure observations. We had trouble adjusting the models to account for this imbalance, and ultimately decided to randomly pick an equal amount of observations as the impure class from the sample of pure observations. Even though this balancing meant we were losing a large number of observations, the modelling results ultimately improved.

We decided to evaluate the classification methods using precision and recall[16] for the 'impure' observations as well as general accuracy measures. Recall measures the proportion of correctly classified cases of a specific against all actually observed cases of that class, whereas precision measures the proportion of correctly classified cases over all cases classified as that particular class by the model.

The XGBoost classifier especially struggled with the class imbalance, and ultimately did not perform very well. In particular, the algorithm performs fairly poorly when attempting to classify impure observations. The low recall and precision values of 0.11 and 0.34 for the impure label are thus not surprising. In contrast, Logistic Regression proved to be much more stable overall. Figure 4.6 shows the resulting confusion matrix after testing set prediction. Although it certainly presents an improvement over XGBoost, it is still fairly far away from a model that can be used in production, with values of 0.43 and 0.42 for recall and precision. The overall accuracy score of 0.69 is deceiving, as it is heavily skewed by the imbalance of the testing set which was not adjusted for class size.



22 Figure 4.6: Confusion Matrix for the Logistic Regression model predictions

4.6. Results and Possible Improvements

Our final plan was to combine the high frequency interval predictions on an hourly basis, and use the combined prediction to ultimately trigger the alerts. Yet, our models were ultimately too inaccurate to achieve any meaningful results even when combined in that way. Overall, the different measurement frequencies, seeming lack of relevant features and generally poor data quality and documentation severely limited us in our attempts. In order to build an accurate and helpful model, the dataset has to be severely improved in terms of quality, documentation and extensiveness in terms of features and observations. It is also important to note that while the dataset contained more than 700000 observations, the actual amount of observations for Silica Concentrate was only about 4000, with only 290 distinct values after removing interpolated hours. Thus, the size of the dataset might seem large, but seems to contain very little relevant information. In terms of modelling approaches, we also discussed other approaches such as more time series related methods or using hourly patterns as features, yet the quality of existing data

[16] (Powers, 2007)

and lack of continuity in the dataset ultimately deterred us from any attempts. While we have not produced a model that is ready to be used in production, we think our approach is still replicable with a better dataset and our feedback on the dataset is valuable in order to provide higher quality datasets in the future.

The link to the Kaggle code for this section can be found in Appendix 7.

5. Conclusion

This report has demonstrated several use cases and supply chain scenarios in which analytics can be used to improve general business strategy as well as parts of the supply chain. For the Olist dataset, our team leveraged association mining to build a model which can be used in marketing campaigns as a product recommendation system based on sales data. Both the cashew nuts and iron ore production cases present clear supply chain scenarios and analytics solution outlines, but unfortunately both presented shortcomings of the provided datasets in terms of structure, quantity and quality which stopped us from producing an accurate solution. Furthermore, although the dataset used for market basket analysis was well structured, the solution also required a workaround due to row overflow issues. Therefore, we provided specific suggestions for improvement given the supply chain context. Although we could not provide a fully built solution, this process allowed us to better understand the necessary standards for datasets in order to enable analytics.

Bibliography

Agrawal, R. & Srikant, R., 1994. Fast Algorithms for Mining Assocation Rules. *Proceedings of the 20th VLDB Conference Santiago, Chile,* pp. 487-499.

Blattberg, R. C., Kim, B.-D. & Neslin, S. A., 2008. *Market Basket Analysis. In: Database Marketing. International Series in Quantitative Marketing.* 18 ed. New York, NY: Springer.

Carnein, M. & Trautmann, H., 2019. Customer Segmentation Based on Transactional Data Using Stream Clustering. *PAKDD 2019: Advances in Knowledge Discovery and Data Mining,* pp. 280-292.

Chen, T., n.d. *Introduction to Boosted Trees.* [Online]
Available at: https://xgboost.readthedocs.io/en/latest/tutorials/model.html
[Accessed 10 November 2019].

Hahsler, M., 2005. Introduction to arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software.*

Hoerl, A. E. & Kennard, R. W., 2000. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 42(1), pp. 80-86.

Islek, I. & Ögüdücü, S. G., 2015. A retail demand forecasting model based on data mining techniques. *IEEE 24th International Symposium on Industrial Electronics (ISIE),* pp. 55-60.

Johnson, T., 2019. *tinuiti.* [Online]
Available at: https://tinuiti.com/blog/ecommerce/supply-chain-optimization/
[Accessed 9 November 2019].

Nanncy, C., 2017. *Supplier Segmentation – The First Step of an Effective SRM Programme.* [Online]
Available at: https://spendmatters.com/uk/supplier-segmentation-first-step-effective-srm-program/

Omayma A.Nada, H. A. W. H., 2006. Quality prediction in manufacturing system design. *Journal of Manufacturing Systems,* 25(3), pp. 152-171.

ORTEC, 2019. *Demand Forecasting and Order Generation.* [Online]
Available at: https://ortec.com/en/dictionary/demand-forecasting-and-order-generation
[Accessed 9 November 2019].

Powers, D. M. W., 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,* Adelaide: Technical Report SIE-07-001.

Rao, B., 1998. *4. INTEGRATED PRODUCTION PRACTICES OF CASHEW IN INDIA.* [Online]
Available at: http://www.fao.org/3/ac451e/ac451e04.htm
[Accessed 22 November 2019].

Santos, I., Nieves, J., Penya, K. Y. & Bringas, G. P., n.d. *Optimising Machine-Learning-Based Fault,* s.l.: Deusto Technology Foundation.

Appendix

Appendix 5: Iron Ore Attribute Description

The link to the dataset can be found below:

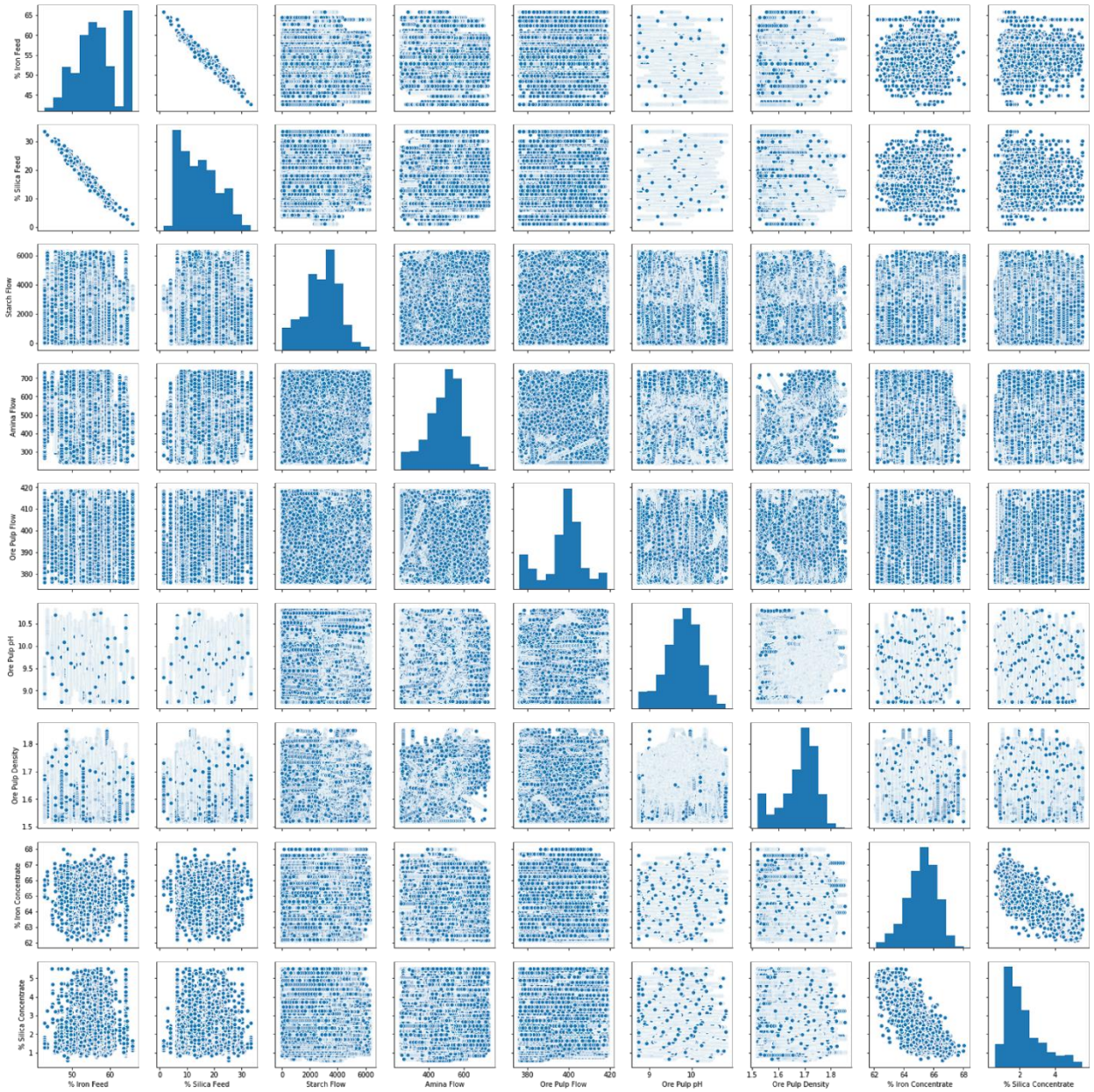
https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process

The dataset columns are shown below with the column descriptions:

- Date: The date and timestamp for the manufacturing
- % Iron Feed: The percentage of Iron which is inserted into the flotation cells and are normally fetched from the Iron ore.
- % Silica Feed: The percentage of silica which is fed to the flotation cells that comes from the Iron ore and it is the impurity for this procedure.
- Starch Flow: The flow of starch in the flotation cells, measured in m^3/h.
- Amina Flow: The flow of amina in the flotation cells, measured in m^3/h.
- Ore Pulp Flow: The flow of Ore pulp during the iron ore production procedure
- Ore Pulp pH: The pH monitored on a scale from 0 to 14
- Ore Pulp Density: Density of the mixture on a scale from 1 to 3, measured in kg/cm^3
- Flotation Column Air Flow (01-07): This field measures the air flow that the flotation cell is provided during the procedure, measured in Nm^3/h.
- Flotation Column Level (01-07): This field measures the froth level that the flotation cell is provided during the procedure, measured in millimeters(mm).
- % Iron Concentrate: This is the percentage of Iron which represents how much iron is the end result of the flotation process. It is normally a lab measurement and represented as a percentage from 0 to 100%.
- % Silica Concentrate: This is the percentage of silica which points how much silica is there as the end result of the flotation process. It is also a lab measurement and represented as a percentage from 0 to 100%.

Appendix 6: Pairplot of Iron Ore Variable Correlations

Paired Scatterplots of % Iron Feed, % Silica Feed, Starch Flow, Amina Flow, Ore Pulp Flow, Ore Pulp ph, Ore Pulp Density, % Iron Concentrate and % Silica Concentrate. The diagonal features histograms of the respective variable.



Appendix 7: Kaggle Link to Iron Ore Production Code

The link to the Kaggle Kernel used to analyse the iron ore production dataset is listed below. The code present in the link is used for the report.

https://www.kaggle.com/mkoerner1/iron-mining-production-prediction