Georges ESSOH

DSC 680: Milestone 2: White Paper

Date:05/08/2022

**Is it possible to forecast employee performance using analytics??**

**Business Problem**

Many industries measure the performance of their employees as a team or individually as their key performance index (KPI). There are a couple of ways this should be possible. One way is through amassing information to oversee work processes and lift efficiency. Worker information can show trouble spots and efficiency spikes every day, and this information gets better with time. Utilizing a performance management system to collect this data can help businesses predict future employee performance. More information can be utilized to assemble baselines of where workers should be at which stages in their profession or current assignment.

**Background/History**

Employees are critical to a company's success. Businesses that recognize this are worried about the output and productivity of their employees. Productivity has a compounding impact at multiple levels in the workplace, which means that strong productivity at one level of the organization leads to higher productivity at higher levels. As a result, in any firm, analyzing employee performance is a must. An employee's performance cannot be traced to a fixed quality. Different people have various skillsets and behavioural characteristics.

**Data Explanation**

The dataset we chose was obtained on UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees) to address this problem. The dataset contains fifteen (15) features and about 1197 instances or rows, and here is the definition of each feature.

**01- date:** Date in MM-DD-YYYY **02- day:** Day of the Week **03- quarter:** A portion of the month. A month was divided into four quarters **04- department:** Associated department with the instance **05- team\_no:** Associated team number with the instance **06- no\_of\_workers:** Number of workers in each team **07- no\_of\_style\_change:** Number of changes in the style of a particular product **08- targeted\_productivity:** Targeted productivity set by the Authority for each team for each day.  
**09- smv:** Standard Minute Value is the allocated time for a task **10- wip:** Work in progress. Includes the number of unfinished items for products **11- over\_time:** Represents the amount of overtime by each team in minutes **12- incentive**: Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action. **13- idle\_time:** The amount of time when the production was interrupted due to several reasons  
**14- idle\_men:** The number of workers who were idle due to production interruption **15- actual\_productivity:** The actual % of productivity that the workers delivered ranges from 0-1. **actual\_productivity** is our target feature.

**Methods**

Two methods are considered the first is to approach or view this as a Time Series Analysis, and the second is to consider this a regression problem. Each of its methods has advantages and disadvantages.

**Time Series Analysis Approach:**

Time series analysis is a statistical technique that deals with trend analysis or time series data. Time series data means data is in a series of periods or intervals, and in this case, the time interval is the day. Time series is about predicting the future values based on previously observed values.

This method is fast and not very greedy from a computational point of view, and its only disadvantage is that it is not very accurate. Despite its lack of precision, it has the advantage of providing a confidence interval within which most, if not all, future values ​​will fall.

We will select two columns from our dataset, namely the *'date'* column and the *actual productivity* column. Because we have two variables, so we have the case of a univariate time series forecasting.

Before the modelling phase, it will be important to prepare the data by Removing factors like trend and seasonality to make the data suitable for the model. The model that will be used is the most popular and is called ARIMA and works very well in the context of univariate time series forecasting.

**Regression approach:**

We have identified three main steps necessary before generating our final predictive model: Exploration Data Analysis, Feature/variable selection, and the consolidation of the two previous steps.

• **Step I:** We will be conducting an exploratory data analysis (EDA) on our dataset. The dataset contains seven (07) different variables that could contribute to our predictive model. At this level, gaining insights into the data is necessary, and it is crucial to identify and deal with missing values and outliers at this level. Visualization will be handy insofar as it will allow us to understand how each variable is distributed. The type of distribution predetermines the type of model. Since most models assume that the data distribution is normal, if otherwise, the distributions should be normalized. At this level, it would also be necessary to proceed to the encoding of certain categorical variables for the sake of the next step, which is the selection of variables, because most algorithms take continuous values as input. Renaming the columns for the sake of simplicity is another thing that will be done.

• **Step II:** At this stage, the focus will be on feature/variable selection from the 15 variables within the dataset after encoding. Several methods such as filtering by variance, Feature selection by correlation, and Feature Selection Using a Wrapper will be tested. The advantage of some of these methods is that they generate top feature data frames. The overall feature score was then determined, which provided the final feature rankings.

• **Step III:** Step I and Step II results are combined at this stage. The final features that will be used in our initial predictive models build. The models will then be run using cross-validation, the summary of results generated, and discussions will follow.

**Analysis**

**Time series Analysis:**

The graph below is a line graph representing the current productivity overtime for January 1, 2015, to March 31, 2015, i.e., three months which is equivalent to a quarter of the year.

Graphical user interface

Description automatically generated

The next step was to determine whether the time series was stationary. If it is non-stationary, we should be differencing the series (once or more) to remove any trend, seasonality, and white noise. Forecasting a stationary series is relatively easy, and the forecasts are more reliable.

Chart, line chart

Description automatically generated

Rolling averages are useful for finding long-term trends otherwise disguised by occasional fluctuations. As we can see, no trend particularly stands out, indicating that the series is stationary. The mathematical confirmation comes from the following table.

|  |  |
| --- | --- |
| ADF Statistic | -4.089394972220049 |
| n\_lags | 20 |
| p-value | 0.0010083639525049711 |
| Critical Values | |
| 1% | -3.4359228156852093 |
| 5% | -2.864000824228055 |
| 10% | -2.5680801943923828 |

With P-value <<0.05 and ADF statistic < to all critical values, it is safe to conclude that the series is stationary, we can then move to forecast.

ARIMA is the model chosen, and the results are compiled in the table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | RMSE | R\_SQUARED |
| ARIMA | 0.08132 | 0.018124 | 0.134626 | 0.404218 |

**Regression Approach:**

* **Exploratory Data Analysis:**

The first graph of our appendix (appendix A) presents pair plots (scatter plots) of all numerical variables of our data set and histograms showing the data distribution. The idea is to know the data distribution for each of these variables. So, we can see that "actual productivity," our target variable is slightly left-skewed. At the same time, none of the scatterplots show any clear correlation between variables or very little correlation with the target variable's actual productivity.

**Bivariate analysis**

Some of the independent variables display high correlation. It is the case of the *number of workers* in each team and the standard minute value, which is the time allocated for a task. The variable *number of workers* and *overtime* seem to be positively and highly correlated, which is also the case of the variables *over time* and *standard minute value*. Targeted productivity and actual productivity display a weak correlation meaning the actual productivity barely meets the expected productivity.

Timeline

Description automatically generated

To complete our analysis, let's look at the distribution of our charges. As the diagram below indicates, the load distribution is left-skewed. Here the distribution tells that the mode is 0,8. But also that the majority of data is below 0.8.

Chart, histogram

Description automatically generated

* **Data preprocessing:**

The dataset has three categorical columns: Quarter, Department, and Day. The quarter consists of Quarter1, 2,3,4 and 5 and data analysis showed that Quarter5 had higher actual productivity than the other quarters. For this reason, we have one-hot encoded this column. The department column consists of sewing and finishing. Finishing has been observed to have higher productivity than sewing. This column is also one-hot encoded. The Day column has six segments: Monday, Tuesday, Wednesday, Thursday, Saturday, and Sunday. The data analysis found that Saturday is the y of the week with the most productivity on average, so we one-hot encoded each of these days as a new column, as shown in the graph below**.**

* **Feature selection:**

Several methods such as filtering by variance, Feature selection by correlation, and Feature Selection Using a Wrapper were tested. At the end of the feature selection process, these features are retained for model buildings***: 'targeted\_productivity','smv','incentive','idle\_men', 'no\_of\_workers', 'finishing', 'sewing'.***

* **Validation process:**

Before modelling, it is important to standardize the results validation process to ensure that it is the same for each of the models tested and that the same data must be used in each model. For this, we have chosen cross-validation. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it. It has a single parameter called 'k,' which indicates the number of groups that the data would be split into. Here we train the model on k-1 datasets and test on the kth dataset, and this process repeats till the k value we set, and in our case, we selected the k value as 10

* **Model Building:**

With the 'actual\_productivity' value to be predicted being continuous, we are in a classic case of regression. We have chosen three powerful models: Linear Regression, KNeighbors regression, and XGboost regression. The results are compiled in the table below.

|  |  |  |
| --- | --- | --- |
| Models | R squared | RSME |
| Linear regression | 0.23 | 0.15 |
| KNeighbors regression | 0.41 | 0.13 |
| XGboost regression | 0.47 | 0.13 |

**Conclusion**

The analysis we have just completed suggests that several factors can affect productivity. That some days, a quarter has higher productivity than others on average. The initial results are not great, and there are several ways this model can be improved to guarantee decent results if moved to a production environment.

The results of the time series analysis are not very different from that of the regression. This indicates that we can dwell on the time series analysis if we want to understand ​​the fluctuation interval of productivity quickly. It gives a confidence interval of 95% that can be used to fix the next productivity targets for the coming months.

The average results of our time series analysis and regression analysis are evidence that other factors must be considered when trying to predict employee productivity. Both approaches only explained 40% of the variance in the actual productivity.

**Assumptions**

Part of the assumptions made is to ensure the integrity of the information collected and the continuity since we have a time series. We understand the sensitivity of collecting such data, but we must ensure the quality of the data collection process.

**Limitations**

One of the main limitations we face is the size of our dataset. Indeed, the more data we have, the better it is for our models.

**Challenges/Issues**

This is both a Time series and regression problem, and one obvious issue is how to deal with wildly different results from using different approaches for feature selection. I will need to develop a plan to deal with such irregularities that might lead me away from the focus of this project. Another issue would be adequately encoding the categorical features, which is key to success.

**Recommendations**

There are several ways this model can be improved for future use because nothing guarantees that it will perform if moved to a production environment. Scalability testing would be required to see if the model can accurately predict employees' performance from new data. Automatic processes can be put into place to blend the data without manual work. This includes data collection as well, by using methods such as web scraping to automatically acquire the data and automatic correlation testing to determine the best variables from new data as well.

**Implementation Plan**

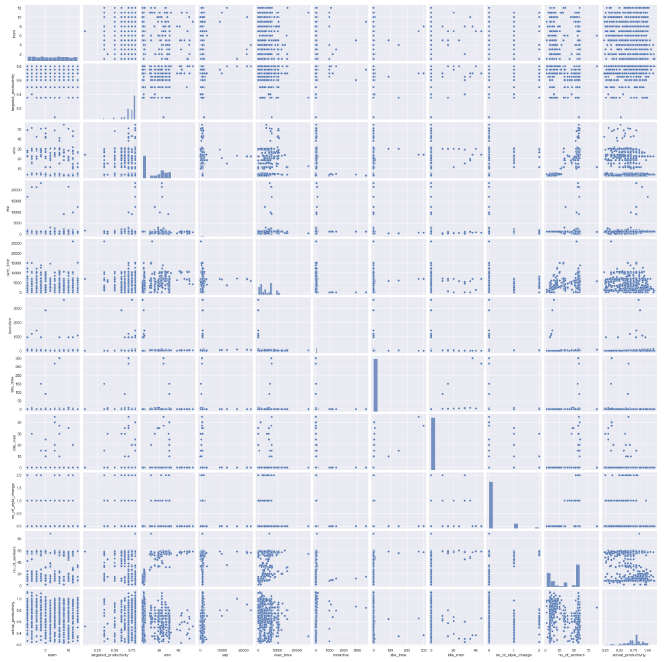
The best and fastest implementation strategy is to implement an application that will generate web pages for the managers and supervisors. This application will stream directly live data and create an estimate of the employees' performance per shift, teams, and days to come.

**Ethical considerations**

It is always difficult to predict an employee's performance because many factors come into play, including personal factors. These personal factors cannot be ignored at the risk of dehumanizing employees and seeing them as mere machines.

**Appendix**

Appendix A



**References**

Prabhakaran, S. (2022, April 4). *Time series analysis in python - A comprehensive guide with examples - ML+*. Machine Learning Plus. Retrieved May 8, 2022, from https://www.machinelearningplus.com/time-series/time-series-analysis-python/

Prabhakaran, S. (2022, April 4). *Augmented dickey-fuller (ADF) test - must read guide - ml+*. Machine Learning Plus. Retrieved May 8, 2022, from https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/

Krishnan, S. (2019, December 20). Variable Selection using Python - Vote based approach. Retrieved from <https://medium.com/@sundarstyles89/variable-selection-using-python-vote-based-approach-faa42da960f0>.

Pickell, D. (2019, May 14). 8 Examples of Industries Using Predictive Analytics Today. g2. Retrieved September 12, 2021, from <https://www.g2.com/articles/predictive-analytics-examples#retail>.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly.