Finding the Biggest Statistical Offenders Causing Delays in Reports Delivery

Muhammad Talha Zahid

Capstone Public Project Summary

June 2022

Background

Financial Markets are really time sensitive; a delay of minutes can have a huge impact on the returns. The company for which this project is done provides investors with financial reports to aid the investment decisions. Delays in report delivery can have multiple outcomes ranging from client dissatisfaction to wrong decisions being made by the client or portfolio managers. Furthermore, as delay cause company increase usage of computational power and under-utilization of resources, necessary steps needed to be taken to increase process efficiency and optimization of resources.

As a part of this project, I'll be taking complete data of all the package reports delivered by the client in last 4 months and selecting the packages which got delayed. In case of delays, analyst have to enter a text in the system explaining the cause for delay. So, I'll be processing those text entries to make binary variables for the factors which causes the delays. I'll be further regressing these independent variables with the delay time and showing how each of the variable contributes to the delaying of the package reports. Based on the results recommendations will be provided to the client to fix those factors.

Data Cleaning

The data of last 4 months of report deliveries was extracted from client database. In total there were 60,000 records of packages to be delayed in the last 4 months. The extracted files were in csv and were directly loaded on Jupyter Notebook for further analysis. The first task was to select the packages which got delayed in the last 4 months. The filtered data had a lot of missing values which after careful assessment of their impact on the target variable were dropped. Additionally, there were multiple duplications in the data which were further removed.

The main task in data cleaning was to process the text entry by the user which would be used to make independent variables. The entered text had punctuation marks, special characters, typos etc and required cleaning. A function was made to remove all the punctuation marks and special characters from the text. The text was mix of upper-case and lower-case characters, so it was converted to lower-case to ensure uniformity in the text.

After exploring the text and understanding the system processes, a list of key words was created to be picked up from the text which would be used as binary variables. A "str.contains" command was used to pick up keywords from the text. After all the keywords were processed, some entries remained which weren't mapped against any of the keywords. On a closer inspection of these records, it was noticed that most of them had typos, such entries were manually mapped against the variables. The remaining user entries were dropped. On exploring the created variables, it was discovered that most of the user text was mapped against more than 1 variable, reason being multiple keywords being present in entries. After discussion with the client, it was decided to make sure that 1 entry is mapped to 1 variable. Some variables seemed to be different names for the same problem which were aggregated. At the end of this step, I had 23 binary variables and 12,000 entries of package delays.

Model Building

To aid the analysis, the created binary variables were categorized into broader categories. The categories were based on the nature of variables. Following are the categories:

1. User Error: Packages delayed because of analyst error; forgetting to close the package before deadline. Such cases were not very common and couldn't be fixed by system or process upgrades.
2. Dependent Reports: For some reports there was an input file which was an output of some independent package. In case the independent package got delayed the dependent package will be delayed as well.
3. Production Factors: Delays caused due to the operational inefficiencies of the system processes which increase overall system runtime of the processes.
4. External Factors: Delays caused due to external factors for instance late files from vendor or client.
5. Internal Factors: Delays cause due to the client's internal system issues like long-running processes.

Technique Used

Ordinary Least Squared (OLS) is the technique used for the regression which is a type of linear least squared method used for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear functions of a set of explanatory variables by the principle of least squares; minimizing the sum of squares of differences between the observed dependent variable in the given dataset and those predicted by the linear function of the independent variable. The OLS estimator is consistent when the regressors are exogenous, when the errors are homoscedastic and uncorrelated. OLS provides minimum variance mean unbiased estimation when the errors have finite variances [1]. *Statsmodels.formula.api* which is a build-in library was used to perform the OLS regressions.

Models

The first model was the most basic one and it took into account just 1 factor of User Error with the delay time of the package. In the second model, I additionally took into account the Dependent Package. In model 3 Production factors, In model 4 Internal factors, and in model 5 External factors. The best model in terms of R-squared and Beta Co-efficient was model 5 which had a R-squared of 6.6% and Beta Co-efficient with higher significance levels. The models failed to establish any direct relationship between our dependent and independent variables in a way that some factors which are known cause of delays for instance QC has negative Beta values meaning in case longer QC the reports would be delivered before the deadline. To check if the presence of some factors are causing others to have negative Beta values, the factors were regressed individually with the delay time to see if the results change, but the relationship remained the same. One possible cause of this can be a high variation in the time of delays caused due to same factors for instance QC have caused a package to be delayed by 1 minute and another time by 1,000 minutes. Another possible cause of this negative relationship can be incorrect user entry which means text entered by analyst at time of delay had nothing to do with the actual cause of delay. As there were no available system logs for the times of processes, it wasn't possible to compare the time of process on the day of delay with the benchmark time. Furthermore, there was no way to check the correctness of entered text.

As a part of improving the results after discussion with the client and checking the distribution of delay times, it was decided to put the filter on packages which were delayed between 0 and 480 minutes. The set filter was applied on the model 5 as it was the most comprehensive one and with the highest R-squared value. The model performed better than the previous models in a way that the R-squared improved to 13.7% and some Beta Co-efficient which were previously negative are now positive with higher significance level as compared to previous models. However, the model still showed negative Beta values for the factors which are known causes of delays.

Another approach to get better results was to select the individual clients based on number of reports delayed for them in the last 4 months. The results improve in terms of higher R-squared values and better Beta Co-efficient but still showed opposite relationship with the factors which are known causes of delays.

Limitations

1. The dataset had very high variance in terms of lateness time for the same reason for instance QC have caused a package to be delayed by minute and another package to be delayed by 1,000 minutes. This high variation in the values might have been the reason of why some variables who were expected to have positive beta values had negative.
2. There is no way to check the sanity of the data, there can be many possible errors in the UserEntry text. The reason of delays might be totally independent of the reasons mentioned in text and there is no way of cross-checking the data. Improved quality data might have resulted in better results.
3. Mapping the variables based on keywords was difficult as most of times a single entry contained more than 1 keyword. Availability of hierarchy of tasks might have made the process easier and efficient.
4. Task of finding causality between variables is still under progress and people are still researching in the field. Due to the on-going nature of the research not a lot of tools are available when you compare this with for example prediction models which have detailed models and machine learning tools available.
5. As the project was done on BLK environment, the available Python notebooks had pre-installed packages and there was limitation in the packages which can be used for instance visualizations were done 'Seaborn' because 'Plotly' wasn't available.

Recommendations

Based on the learnings of the project following are the recommendations for BlackRock:

1. **Improve Quality of Data**: There should be logs of system processes for packages, such details will make analysis easier and provide better quality results. With the available log files, weights can be assigned to every process because every process is unlikely to take same amount of time, so it is not accurate to give them same weightage when calculating the impact on delays.
2. **Broader Categories for User Entry Text**: Increasing the categories of available reasons for delays for analyst to select in case of delays will be helpful in understanding the real cause of delays. Providing training to the analysts to ensure every part of process is clear will also provide better quality results.
3. Based on the models implemented, BLK can work to reduce the time of delays caused by late Client Files, Midweek month-end production, and long running covers. These variables have high significance levels and are the most known causes of delays.
4. An alternate approach was suggested but couldn't have been implemented because of time constraints, Clustering using K-means is an unsupervised learning method in which no labels are given to the learning algorithm, leaving on its own to find structure in its input. . In case of any future projects on this topic it would be a good point to start from.

Conclusion

Based on the regression results of above models, none of the model performed how it was expected to perform. The models kept on giving negative Beta Co-efficient for the variables which are known to increase the delay time of a package report. The reasons for this can be multiple ranging from the fact that

UserEntry might not include the real reason why the package got delayed. There is no way of cross checking the UserEntry because there are no system logs of time of beginning and ending of jobs which run in the production of package reports. In case of the log's availability, it would have been easier to compare the time when process actually ran or finished against the benchmark time of the process. Another possible reason for the results can be the timeline of dataset selected, expanding the timeline of the dataset could have provided better results or displayed stronger relationship between the variables and outcome in form of better Beta Co-efficient values or higher R-squared value. Filtering the lateness time and the Sites improved the results but still failed to establish any concrete relationship between the dependent and independent variables. Although Model 5 performed comparatively better in terms of R-squared and Beta Co-efficient values but still cannot be selected as the results weren't significant enough. In short it can be said that this exercise is inconclusive, and quality of data should be improved before any form of analysis can be carried out.

Citations

1. "Ordinary Least Squares." *Wikipedia*, Wikimedia Foundation, 7 June 2022, https://en.wikipedia.org/wiki/Ordinary_least_squares.