**Summary**

This is a summary of the completive report created by group “RER Sucks”. The main objective of our project is to find out the incident reasons of Paris Metro and RER and use machine learning algorithms to predict breakdowns reasons for a certain line in a certain date.

Although people complain about the Paris Metro and RER frequently, there isn’t any existing available data source. The most important part and difficult part is to create our own dataset for future analyzing.

**Create a new dataset.**

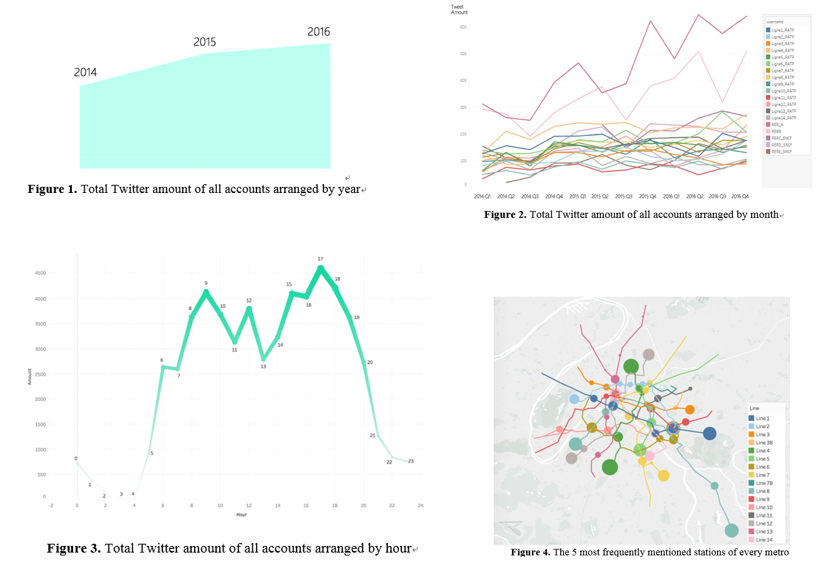
We built a Python web scrawler to scratch all the tweeters sent by 14 Metro lines and 5 RER lines, from 1st Jan.2014 to 21st Oct.2017 and create a 175,044 x 7 data frame consisting of the following information: *username: which line tweeted the tweeter; date: when the twitter was tweeted; retweets: the number of retweets; favorites: the number of favorites; text: the content of the twitter; geo: where the twitter was tweeted; mentions: the twitter mentions what accounts.*

Additionally, we take weather conditions into account and select the following variables to help predict the breakdowns and analyze the reasons: *average temperature; average visibility; average wind speed; rain or not; fog or not; snow or not*.

Besides, we also considered that which arrondissements passed through are also linked to the delay/cancellation of the public transportation.

Finally, we decided to take traffic volume into consideration. We found a data source from SNCF Open Data[[1]](#footnote-1):*Station Name; 2016 Total Voyagers Number.*

**Explore data**

Before we start applying any machine learning algorithm, we do some exploratory analysis first and find several hidden characters of Paris Metro and RER.

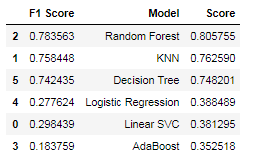
1. The number of breakdowns is growing year by year，which indicates the conditions of Paris underground transportation are getting worse.
2. The occurrence of breakdowns has strong seasonality. The transportation gets worse in bad climates.
3. Breakdowns are more likely to happen at rush hours, such as 9:00 and 17:00, which may cause more troubles.
4. Breakdowns are more likely to occur in the terminals and the joint stations. Among all the stations, “Nation” has the most breakdowns and RER A performed worst comparing to all other lines. In the Appendix of our final report, we listed the top five stations which are most frequently mentioned of each line.
5. The two main reasons leading to breakdowns are abandoned baggage, passenger issues and technical problems.

**Feature engineering**

As we learnt from the Big Data Analytics class, data scientists spend most of their time cleaning and collecting their data. We spent most of time building our own web scrawler and cleaning (text mining) our data, including: normalized all abbreviations and accents, find different characteristics for different twitter accounts, group incident reasons, apply Natural Language Processing techniques to text mining.

Lately, as our primary objective of forecasting the reason of a breakdown of RER/metro lines is a classification problem, we decided to use the following Machine Learning models: *Decision Tree; Random Forest; AdaBoost; KNN; Linear SVC; Logistic Regression.*

**Model Performance**



**Conclusion**

Abandoned baggage is the primary reason for most delay/cancellation of all metro/RER lines. Besides, with a 61.2% accuracy of predicting the incident reason, our model can help RATP and SNCF prevent incidents and improve their service in Paris Region. However, due to the limited time and limited data source, our analysis can be enhanced by a possible cooperation with RATP and SNCF agencies.

1. https://data.ratp.fr/explore/dataset/trafic-annuel-entrant-par-station-du-reseau-ferre-2016/ [↑](#footnote-ref-1)