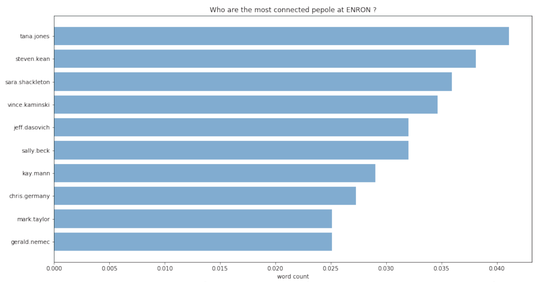
ENRON INVESTIGATION

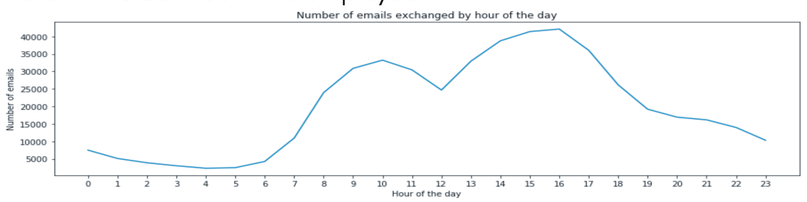
Introduction

In 2000, Enron was one of the largest companies in the United States. Its stock kept soaring higher and higher day by day. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, a significant amount of typically confidential information entered into the public record, including around 500.000 emails and detailed financial data for top executives. The fall of Enron gave us the opportunity to dive into a detailed analysis of emails.

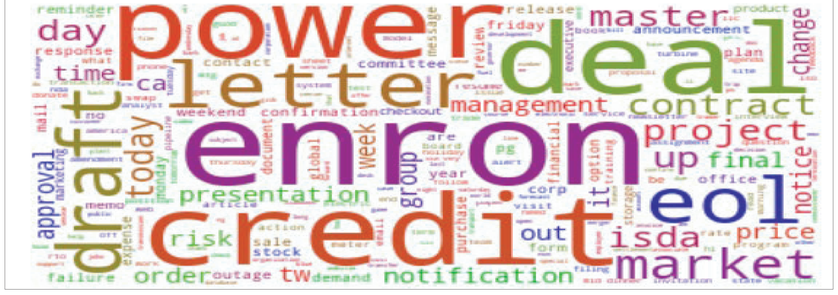
Data Exploration



Initially, we were interested to see the connections between the employees of Enron. We thought that this will give us a great insight about the most influencial people in the firm. We therefore decided to represent the relations between employees as a graph where each node represents an employee and each edge represents a score of connectivity, correlated to the number of each e-mail transmitted between two employees.



Moreover, We can get the exact daily schedule of the employees. When looking at the exchange of e-mail, we can notice that their day starts at 8AM and finish at 6PM with a one hour break at noon. They are quite productive at 10AM and reach their highest productivity at 4PM before people start to leave.



SPAM Detection

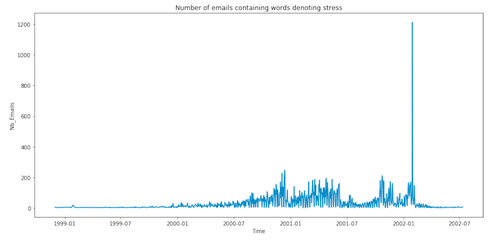
The best way to use a database of emails like this one is to try to create a classifier for SPAM detection. The results we found are very promissing.
We can clearly see that Linear SVM, with a great accuracy of 98.99%, is the best model among the ones we have tested.

	Accuracy
Linear SVM	**0.989948**
SGD Classifier	0.984681
Naive Bayesian Multinomial	0.986802
Logistic Regression	0.982977
Extra Trees	0.973425
Random Forest	0.973754
AdaBoost	0.972014
Extra Forest	0.944377
k-Nearest Neighbors	0.785367

Table 1: Different ML algorithms results.

Fraud Detection

The idea was to build some machine learning models to detect frauds.We therefore started by enriching our dataset with financial data of enron employees which was made public after the scandal.



As we wanted to build a supervised model first, we looked all over the internet for people who reached a settlment with the government, pleaded guilty or testified in exchange of prosecution immunity and labeled them as People of interest.

Aiming at adding more weight to e-mails content in our models, we assumed that the high presence of words denoting fear such as 'trial' or 'fraud' could indicate well if a person was of interest or not so we added a stress score component to our model. While linear classifiers didn't operate well due to unbalancy of data, other models based on trees achieved over 83% of accuracy and 95% of recall which was beyond our expectations regarding the small amount of data at hand.
Since labeled data is expensive we decided to try building an unsupervised model using K-means. It was intuitive for us to run the algorithm using K=2 in order to separate innocent people from guilty ones. Using no prior knowledge of the scandal and despite the very low amount of data and the relative important number of features we got outstanding results. We were able to guarantee that these two people were the two main responsibles of the fraud : Lay Kenneth : founder, CEO and Chairman of Enron Corporation for most of its existence and Jeffrey Skilling : CEO of Enron Corporation during the scandal.

Jeffrey Skilling

Lay Kenneth

Conclusion

The Enron scandal gave us an opportunity to conduct studies on a real life email data-set that would have been very hard to come by because of a lot of regulations and privacy reasons. From data exploration to SPAM/HAM classification and Fraud detection, we discussed what kind of analysis one can do on this type of data. We have had hands on experience on how to collect data from every source imaginable. We experimented with a lot of machine learning algorithms and learned how to correctly choose a model that outperforms every other one facing real challenges such as poorness or unbalancy of data.

The Team

Sami Ben Hassen: sami.benhassen@epfl.ch
Firas Kanoun: Firas.kanoun@epfl.ch
Ali Fessi: ali.fessi@epfl.ch

Github:
https://github.com/fessi12/Project-ADA-2018-Tornada