# Approach Note – Xtreme ML Hack

Author: Sonny Laskar

## About the Problem:

This was one of the best competitions till now on Analytics Vidhya. Moreover, the introduction of submission limits made it even more interesting. This competition was definitely not a good place to start from if you had just started learning ML. My first thought on looking at the problem made me think that it had to be solved as a time-series problem which later seemed to be wrong.

## Modelling Approach:

It was a no-brainer to find that the only data in the Test Set were “future” dates. Did that look odd? Come to reality! The variable to predict was # of contacts/resolutions. After some initial analysis, I decided that this has to be treated as two different problems.

For the modelling purpose, I created daywise-mediumwise-departmentwise aggregated # of contacts and resolutions from 2010 onwards (since we had data on contacts/resolutions only after 2010). For cross validation, I decided to use the last four months. So my first model was built on few “Date” features and it scored **101.X** which was a very good score on Day1.

The next thing that struck me was that holidays should have an impact on these contacts and resolutions. Hence I created a list of holidays in Spain since 2010. Just adding holidays to the above list improved the score to **93.X**. I was feeling good but it didn’t last longer since I saw Mark jumping to **78.X**.

Later, I decided to add “lag” features of # of contacts / resolutions from past. After a few iterations, I decided to keep lag features of 75, 90 and 120 days. With this, the score improved to **64.X** (ahead of Mark ☺). Rohan was around **114.X** so I knew that he was either asleep or solving Soduku. This was on Saturday morning and I was unable to find any additional features that could help me move ahead. So I decided to take a break. By evening, I noticed that Rohan had woken up and was also at around **78.X** (That’s when I guess Mark and Rohan decided to team up).

On Sunday, I added a feature on number of days (percentile) elapsed since last holiday and that added few points. My final 10-bags Xgboost ensemble scored **61.47081** on public leaderboard which I selected for “final submission”. I had around 8 submissions left which I wish I could donate to few folks who were not aware that they could not mark their final submissions if they didn’t upload code ☹.

Huh!

It might sound that the journey of adding features and improving scores was a very smooth line something like this:

But it was not. I tried many things which didn’t work.

Below are few approaches that I tried but didn’t seem to add value:

1. Using # of contracts added/ended data and using them as lagged featured
2. Predicting # of contracts added/ended (time-series and regression) and use them in model (While CV improved but the leaderboard dropped , may be because the predictions of contracts were themselves not correct)
3. Using contacts info for prediction of resolutions and vice-versa
4. Stage 2 meta modelling with predictions
5. Log and Square root transformation of contacts and resolutions
6. Time-series modelling approach instead of regression
7. And many more…

Overall, this was a different problem and really enjoyed solving it. Thanks to Mark and Rohan for a tough fight.