Predicting Unproductive Home Call Doctors

Felipe Fagundes
General Assembly – Final Project

Overview

I work in a Home Care Company in Brazil.

We operate in Brazil for over 2 year. Since them we collected data regarding all the appointments made in patients houses.

Whenever the doctor encounter some issue or problem he could not solve or prescribe something, we need to escalate the issue requesting another doctor with more experience or a specialist to go to the patient residence causing to double our expenses We call this Intercurrence.

This project is to predict when intercurrence may occur so we could prevent to send two doctors to the same house

Data

Started with over 170.000 appointment and 21 Features
Created 9 new features data combining data.

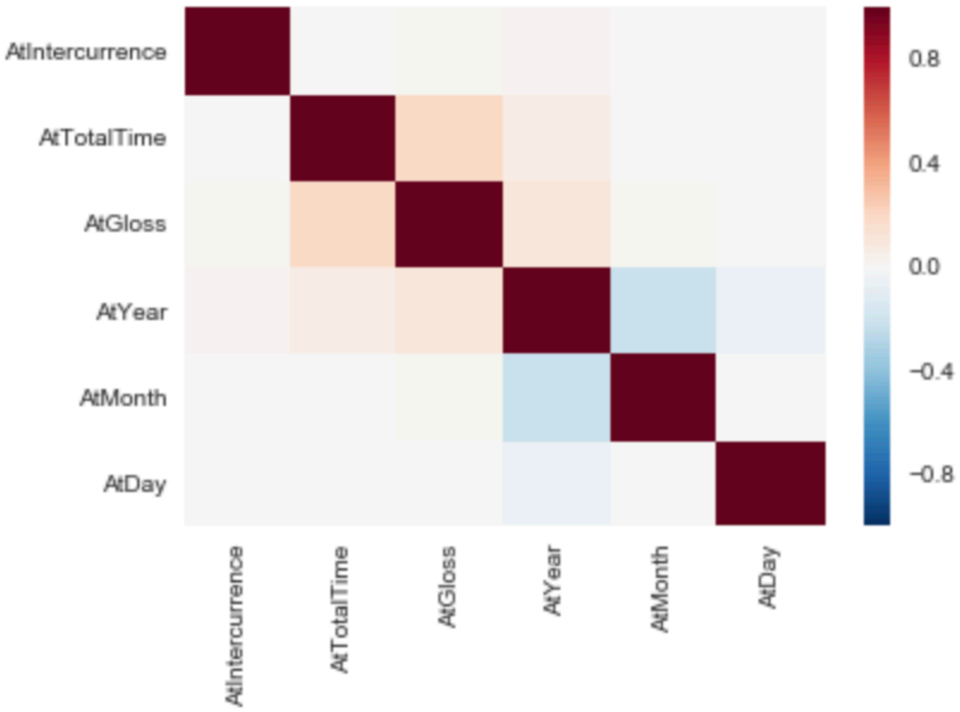
First issue, the data was terrible, too messy, missing a lot of values.

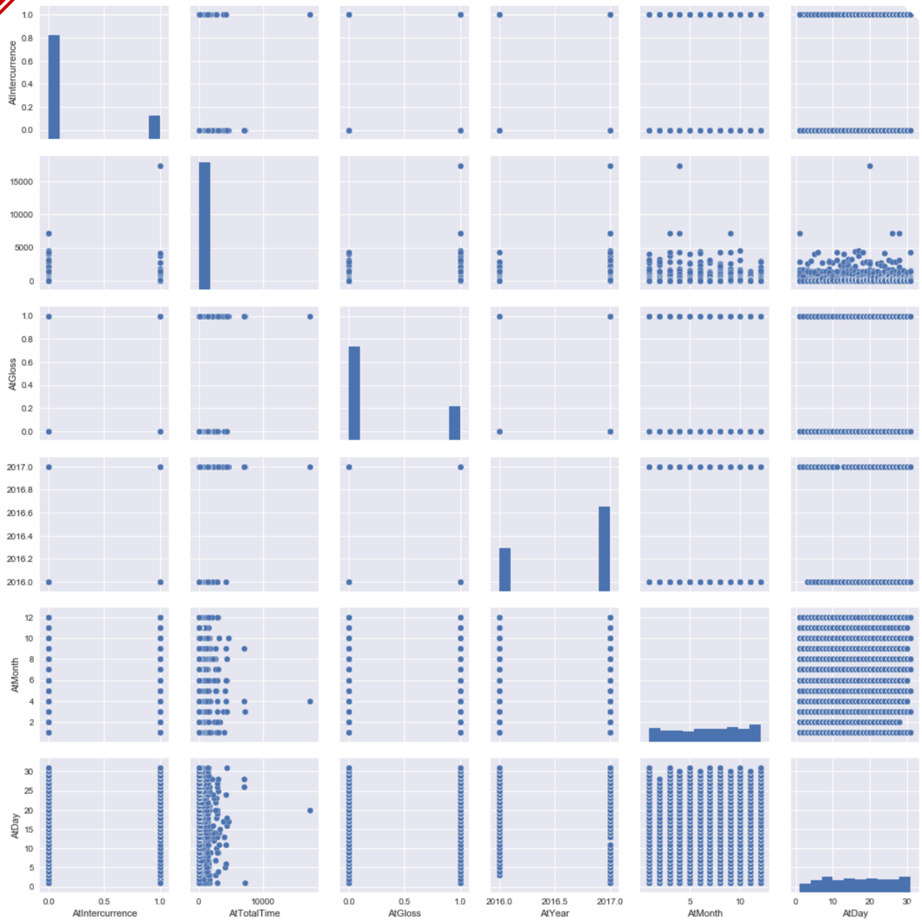
Second issue, was almost impossible to predict intercurrences because more than 95% of the appointment didn’t have intercurrence.

After a good cleanup and re-arranging data I was able to created dummies variable for all categorical values and could start working with some models.

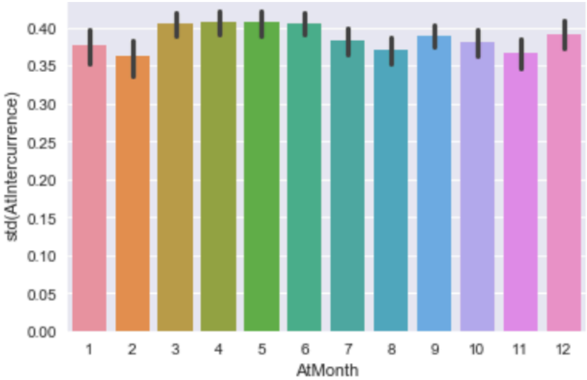
Final data: About 1.500 appointment and 94 Features

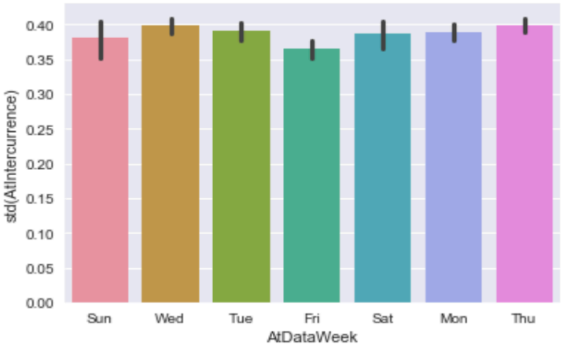
Exploring Data

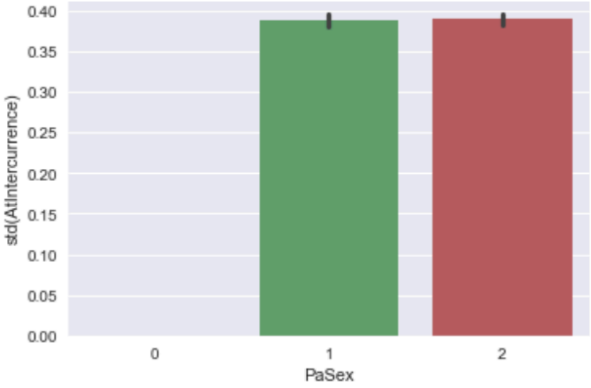


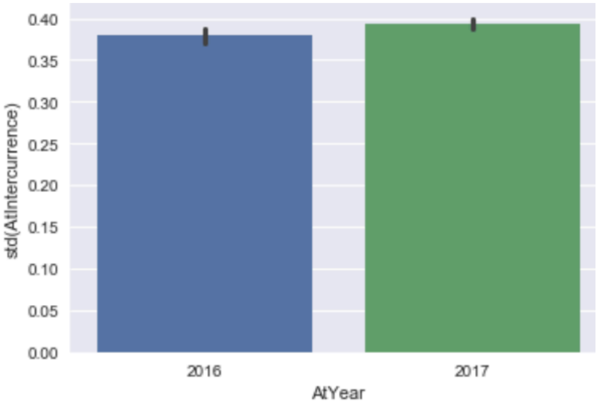


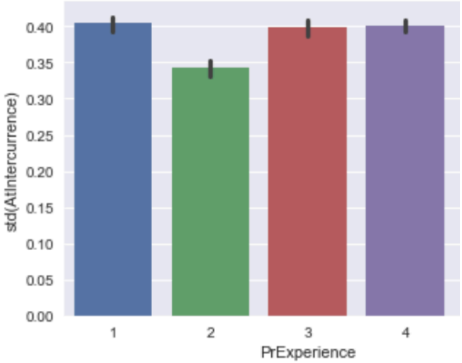
Exploring Data

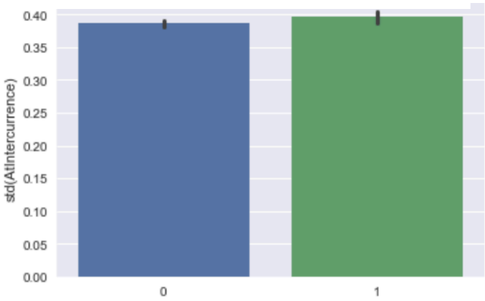


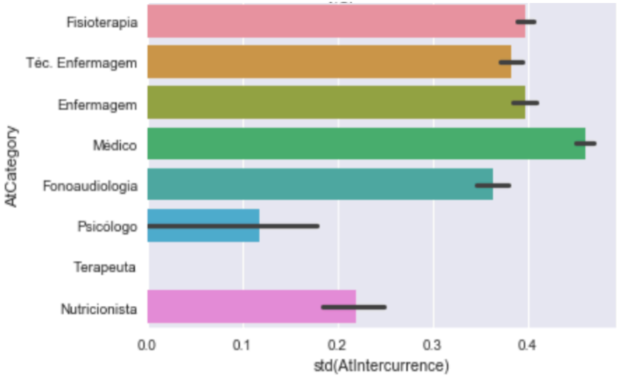




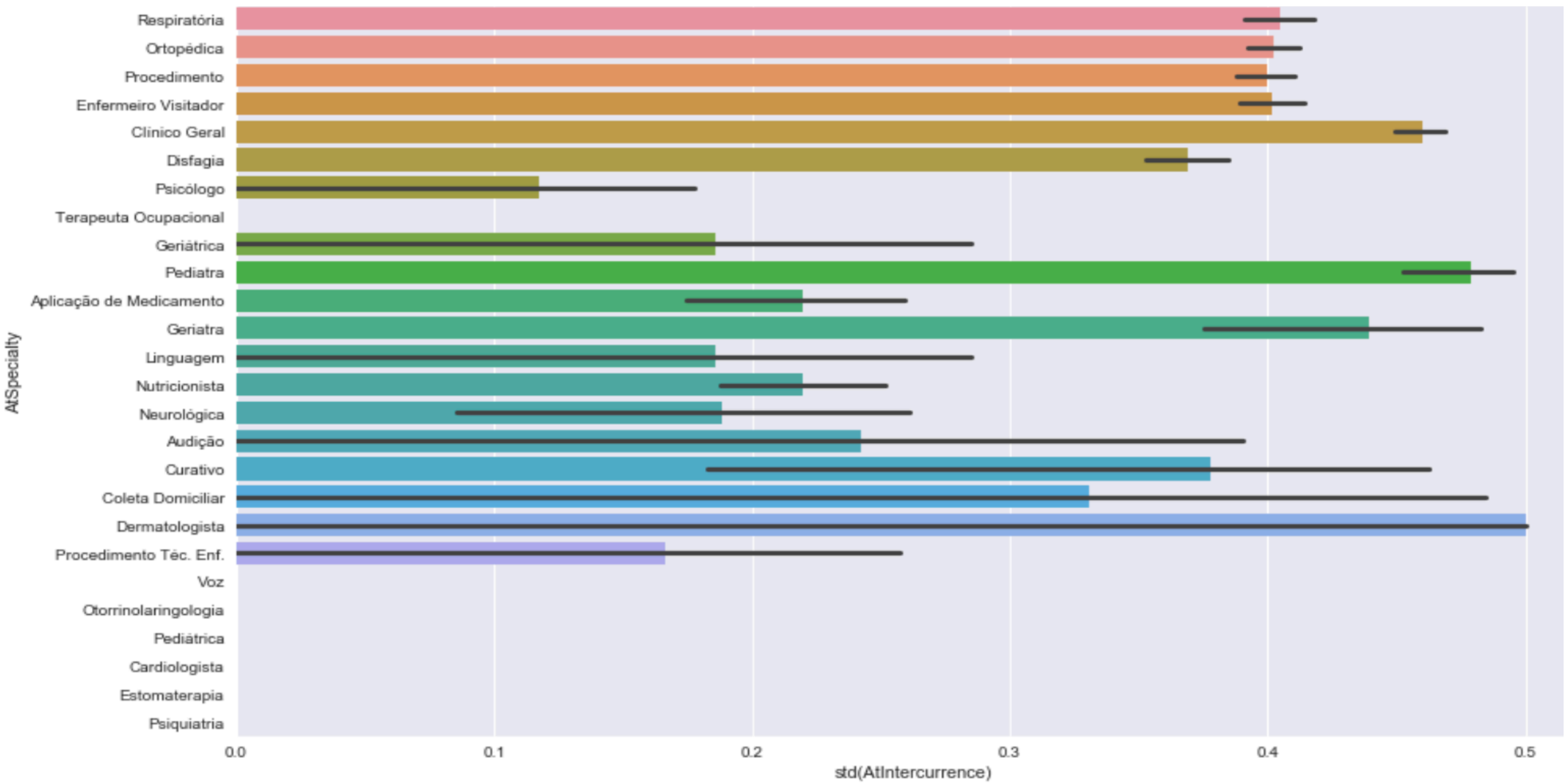




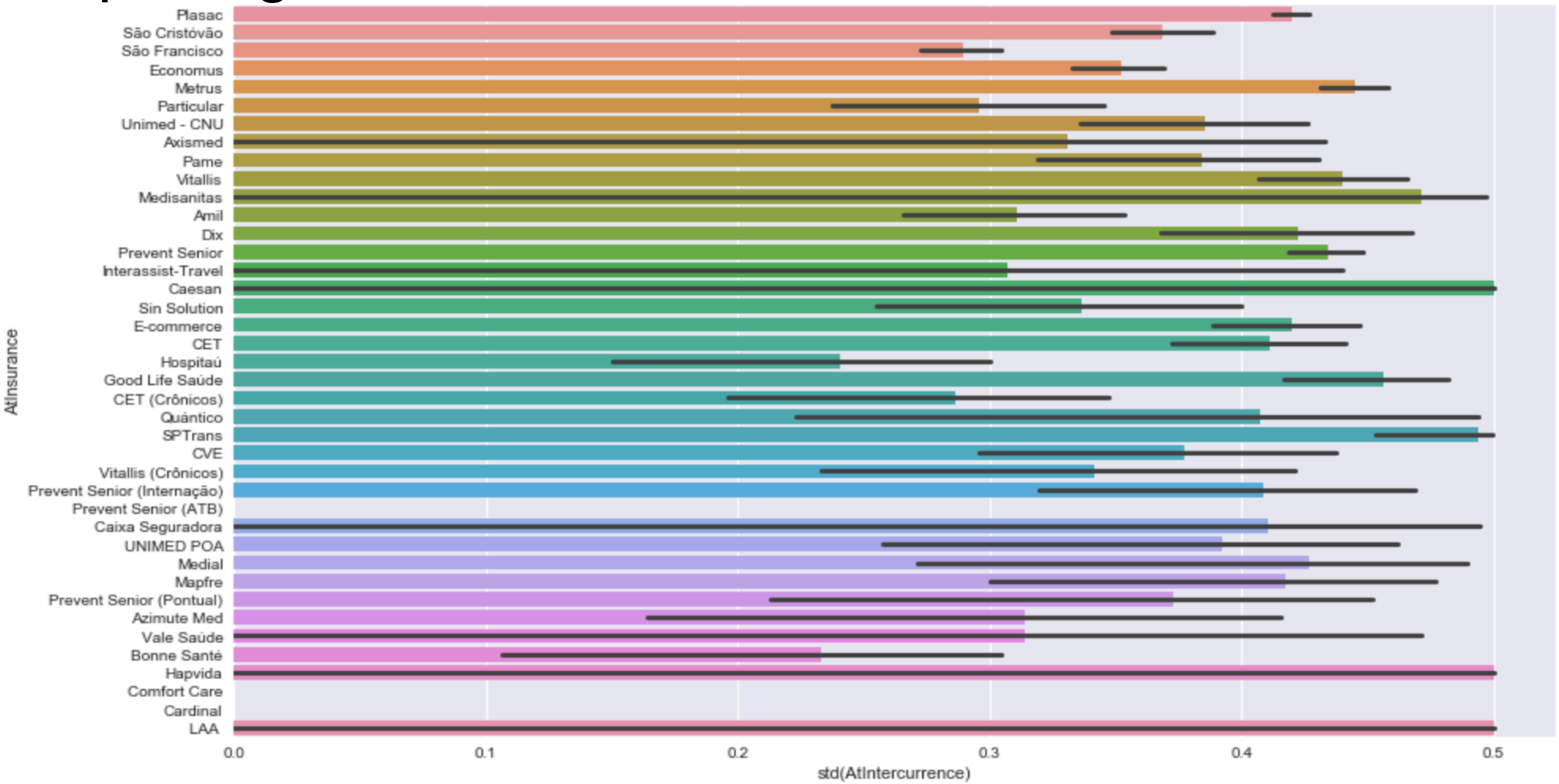




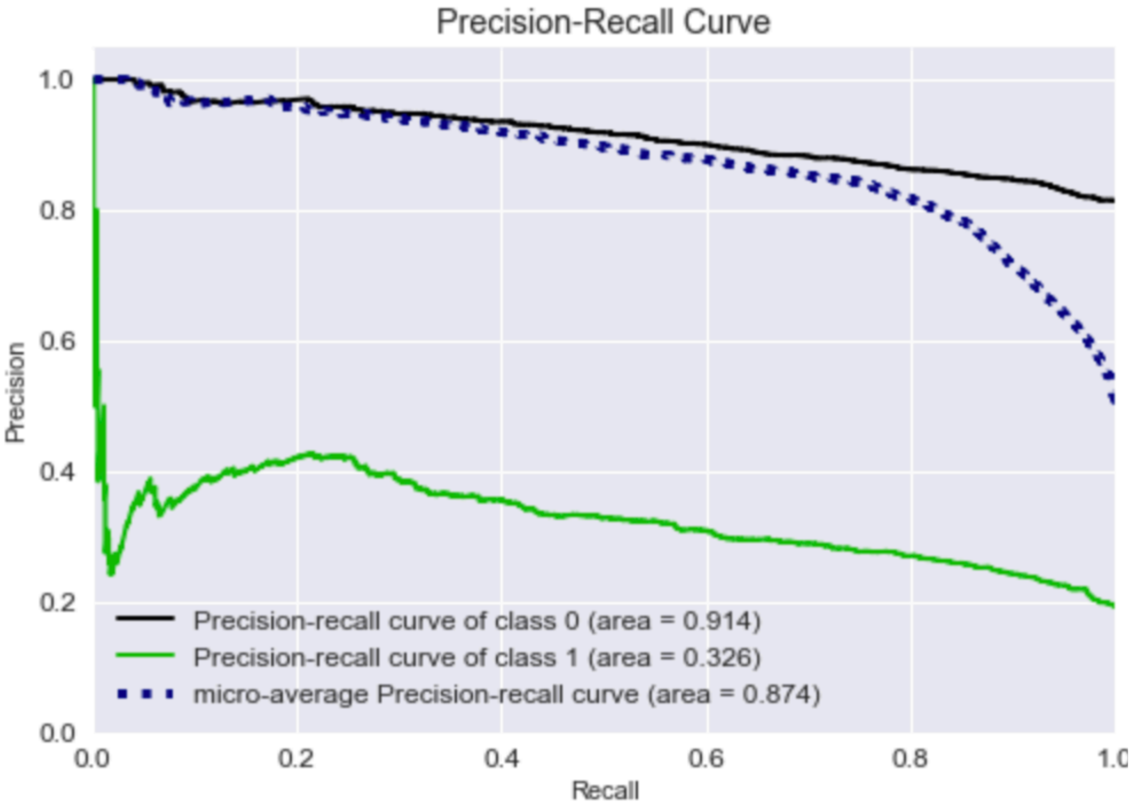
Exploring Data

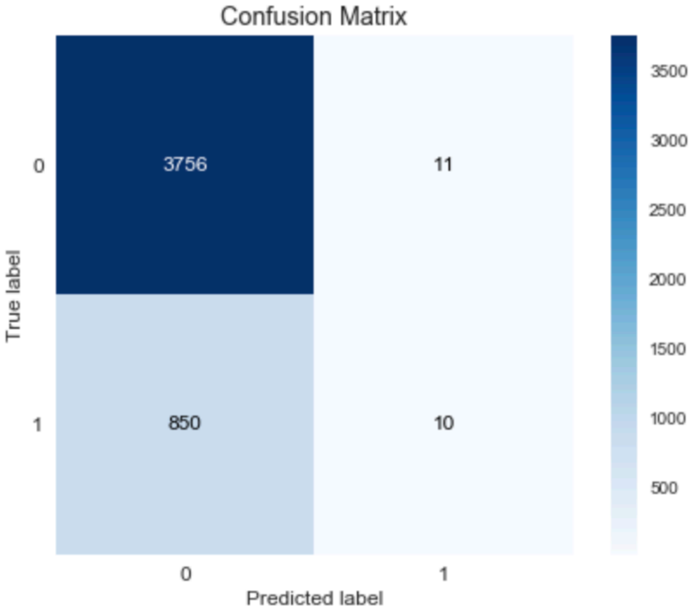


Exploring Data

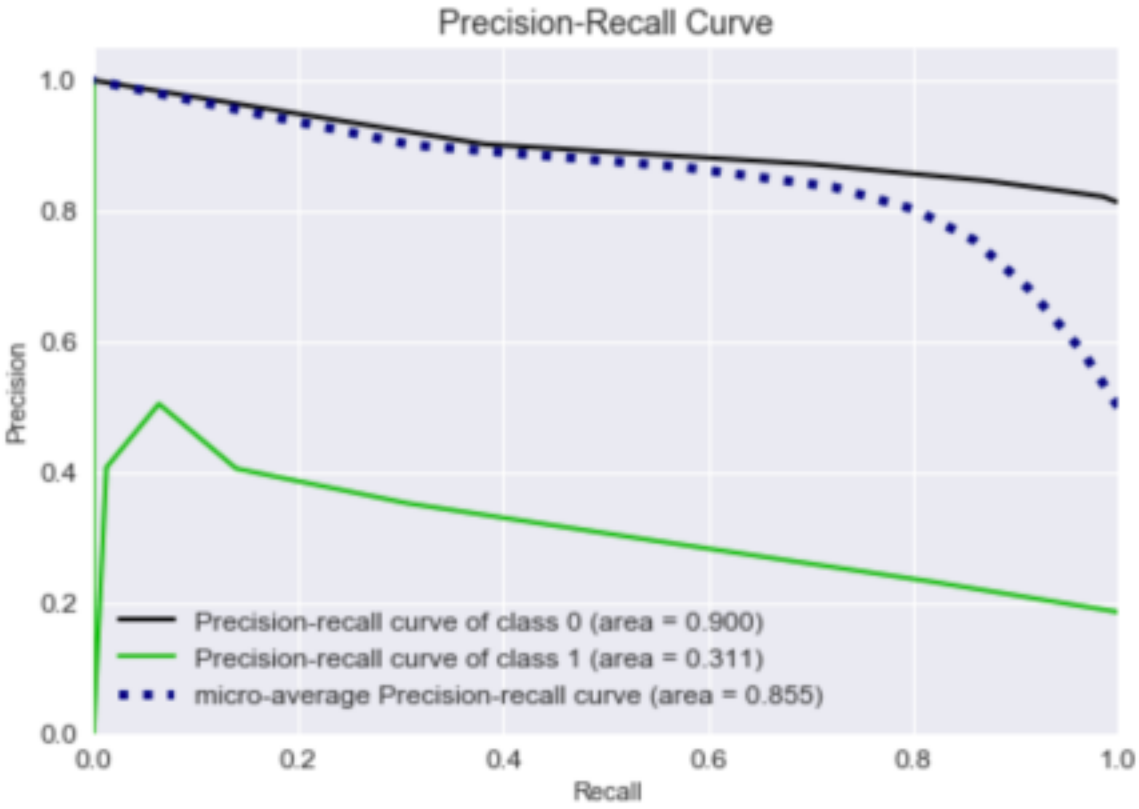


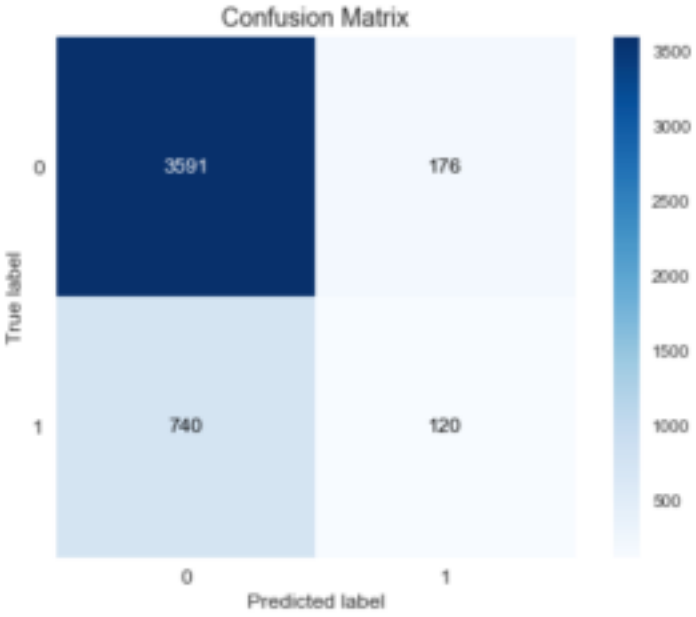
Logistic Regression





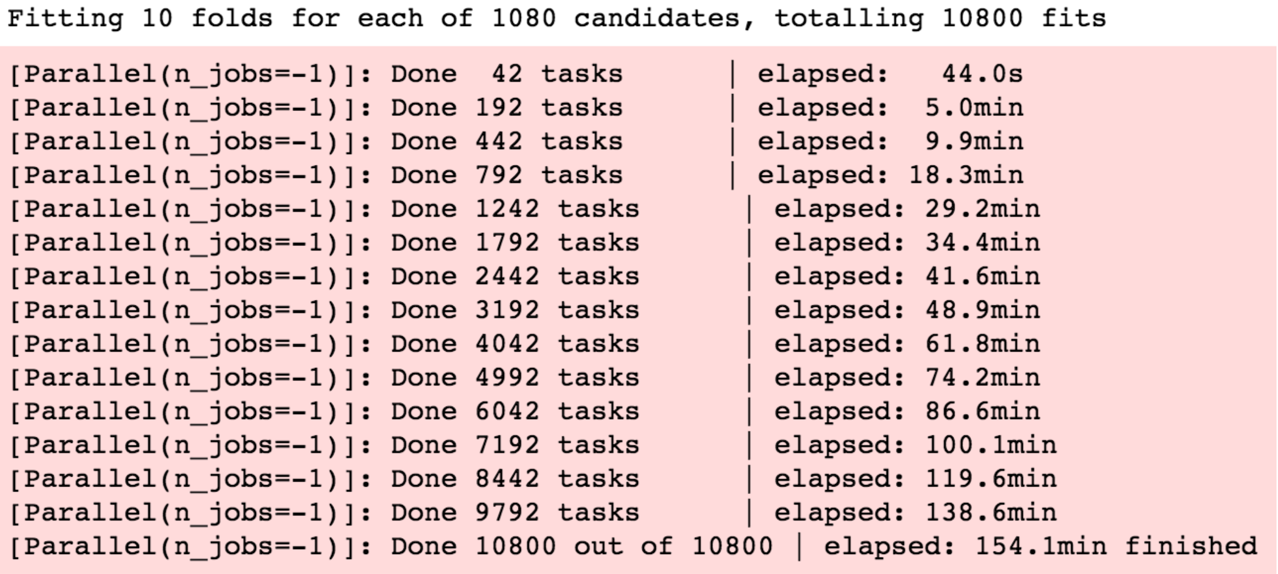
KNN





	0	1
0	3591	176
1	740	120

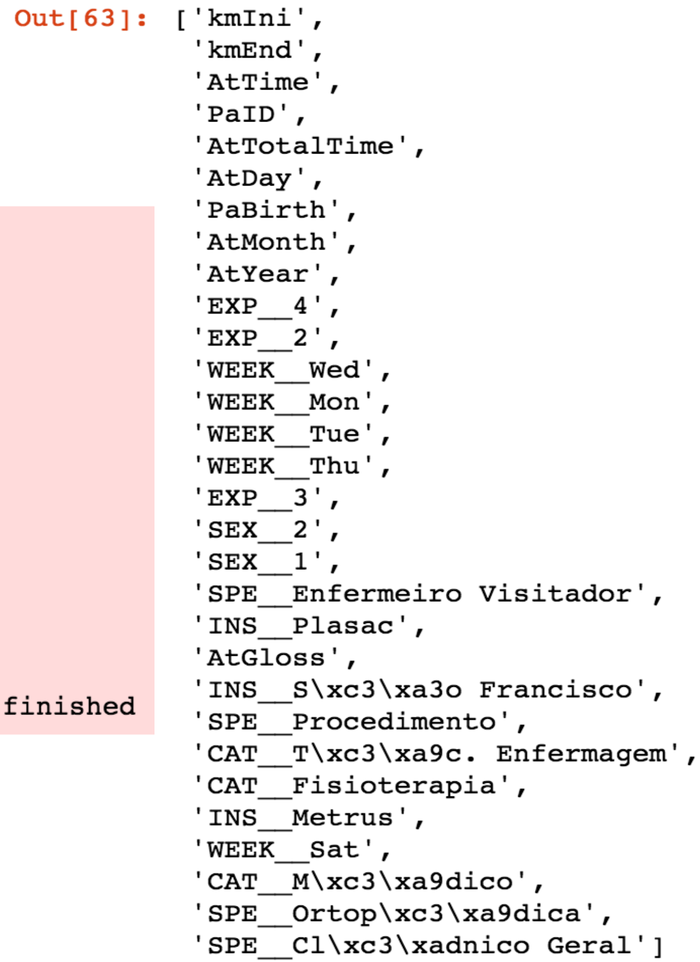
Random Forest



```
Fitting 10 folds for each of 1080 candidates, totalling 10800 fits
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed:   44.0s
[Parallel(n_jobs=-1)]: Done 192 tasks      | elapsed:  5.0min
[Parallel(n_jobs=-1)]: Done 442 tasks      | elapsed:  9.9min
[Parallel(n_jobs=-1)]: Done 792 tasks      | elapsed: 18.3min
[Parallel(n_jobs=-1)]: Done 1242 tasks      | elapsed: 29.2min
[Parallel(n_jobs=-1)]: Done 1792 tasks      | elapsed: 34.4min
[Parallel(n_jobs=-1)]: Done 2442 tasks      | elapsed: 41.6min
[Parallel(n_jobs=-1)]: Done 3192 tasks      | elapsed: 48.9min
[Parallel(n_jobs=-1)]: Done 4042 tasks      | elapsed: 61.8min
[Parallel(n_jobs=-1)]: Done 4992 tasks      | elapsed: 74.2min
[Parallel(n_jobs=-1)]: Done 6042 tasks      | elapsed: 86.6min
[Parallel(n_jobs=-1)]: Done 7192 tasks      | elapsed: 100.1min
[Parallel(n_jobs=-1)]: Done 8442 tasks      | elapsed: 119.6min
[Parallel(n_jobs=-1)]: Done 9792 tasks      | elapsed: 138.6min
[Parallel(n_jobs=-1)]: Done 10800 out of 10800 | elapsed: 154.1min finished
```

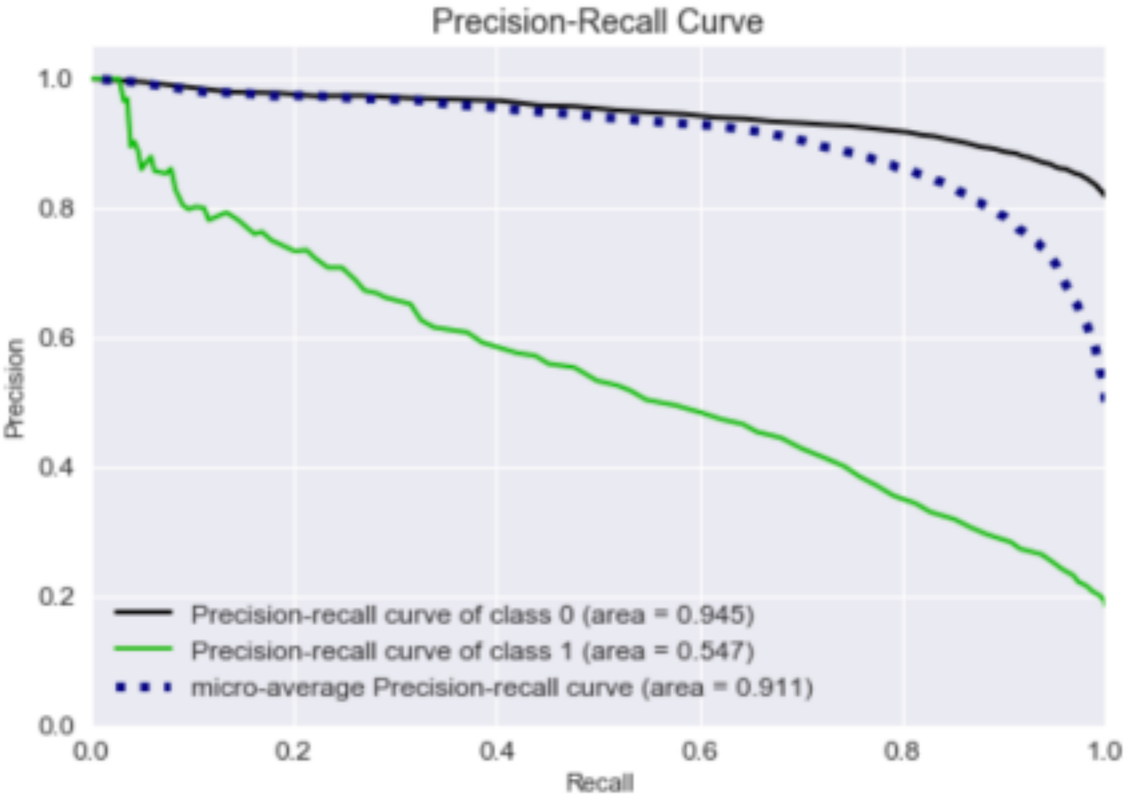
Key hyper-parameter was **class_weight**

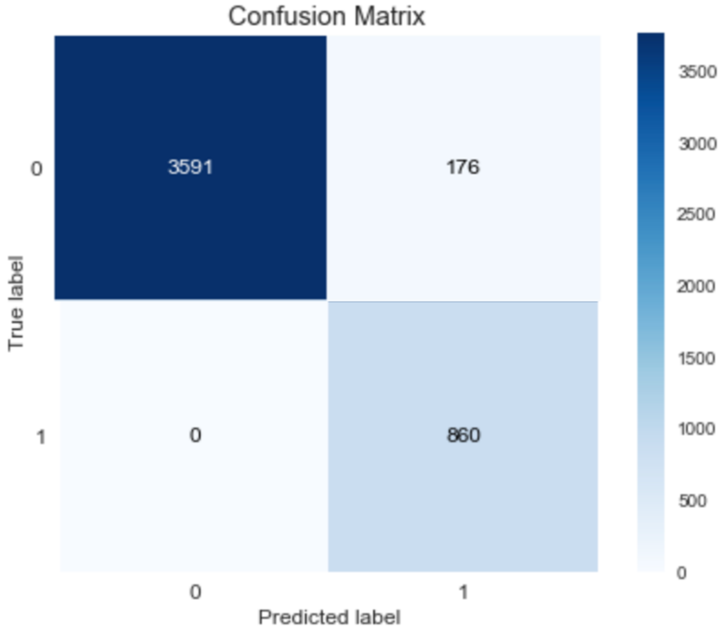
TOP 30 Coefficients



```
Out[63]: ['kmIni',
          'kmEnd',
          'AtTime',
          'PaID',
          'AtTotalTime',
          'AtDay',
          'PaBirth',
          'AtMonth',
          'AtYear',
          'EXP__4',
          'EXP__2',
          'WEEK__Wed',
          'WEEK__Mon',
          'WEEK__Tue',
          'WEEK__Thu',
          'EXP__3',
          'SEX__2',
          'SEX__1',
          'SPE__Enfermeiro Visitador',
          'INS__Plasac',
          'AtGloss',
          'INS__S\xc3\xa3o Francisco',
          'SPE__Procedimento',
          'CAT__T\xc3\xa9c. Enfermagem',
          'CAT__Fisioterapia',
          'INS__Metrus',
          'WEEK__Sat',
          'CAT__M\xc3\xa9dico',
          'SPE__Ortop\xc3\xa9dica',
          'SPE__Cl\xc3\xadnico Geral']
```

Random Forest



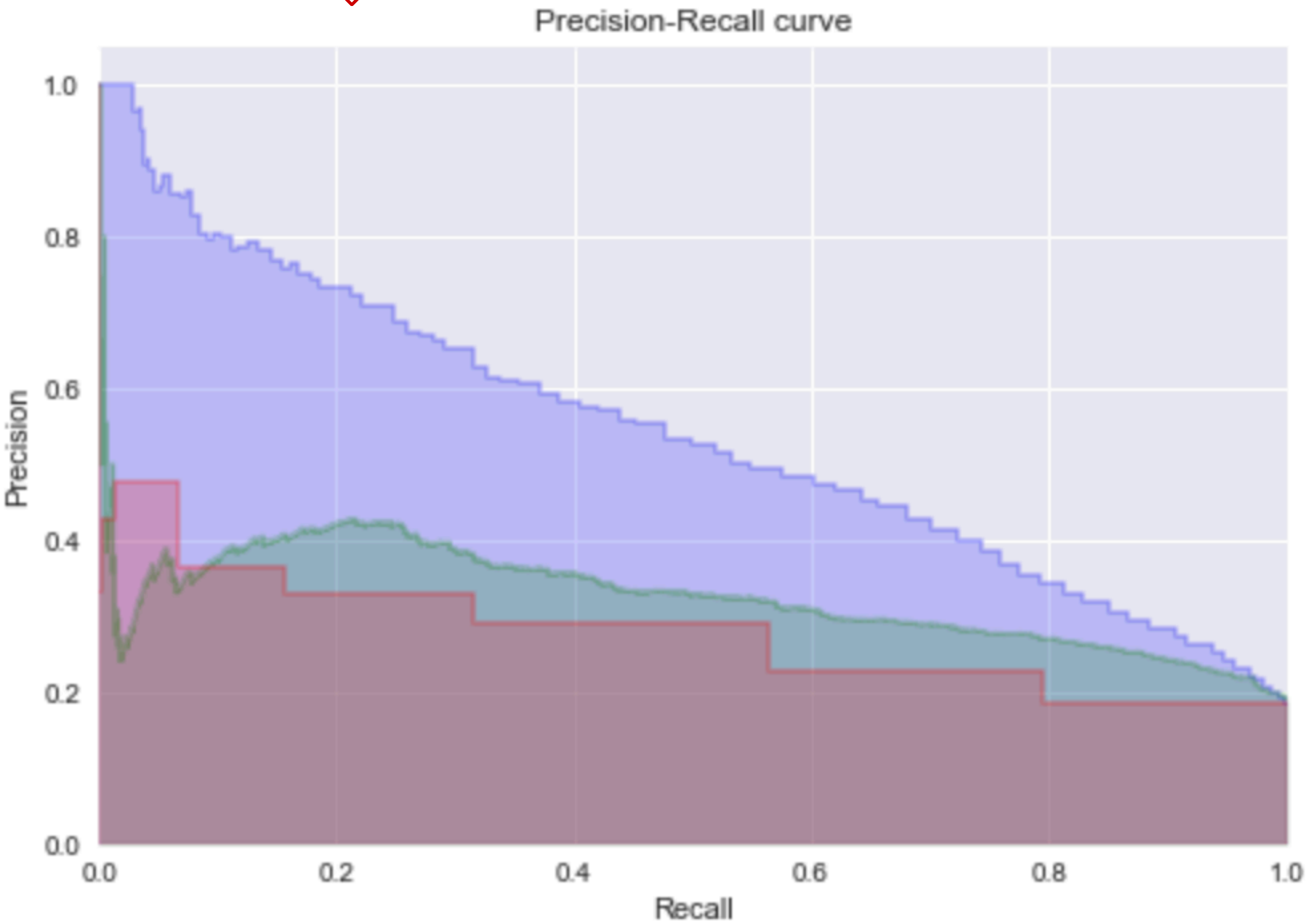


True label \ Predicted label	0	1
0	3591	176
1	0	860

Conclusion

Since my propose is to predict when Intercurrence will happens, I plotted the prediction of the value one (with intercurrence) for all three models.

I could clearly see the blue one (Random Forrest) is so much better than the others and could be improved even further.



Future Work…

- Collect more and different data
- Improve the model with new data
- Improve the current system to collect more accurate data
- Since we have couple of open text fields, use text classification
- Cross information with hospital and insurance data
-
-
-

Thank you…