How Soon Will a Complaint be Resolved?

A CASE STUDY ON NEW YORK CITY 311 CALL

Problem

We accept 311 complaint phone calls 24 x 7 and try to help people.

Satisfied response time?

→ Model the response time

→ (Alternatively) Model the response time category (> 1 day?)

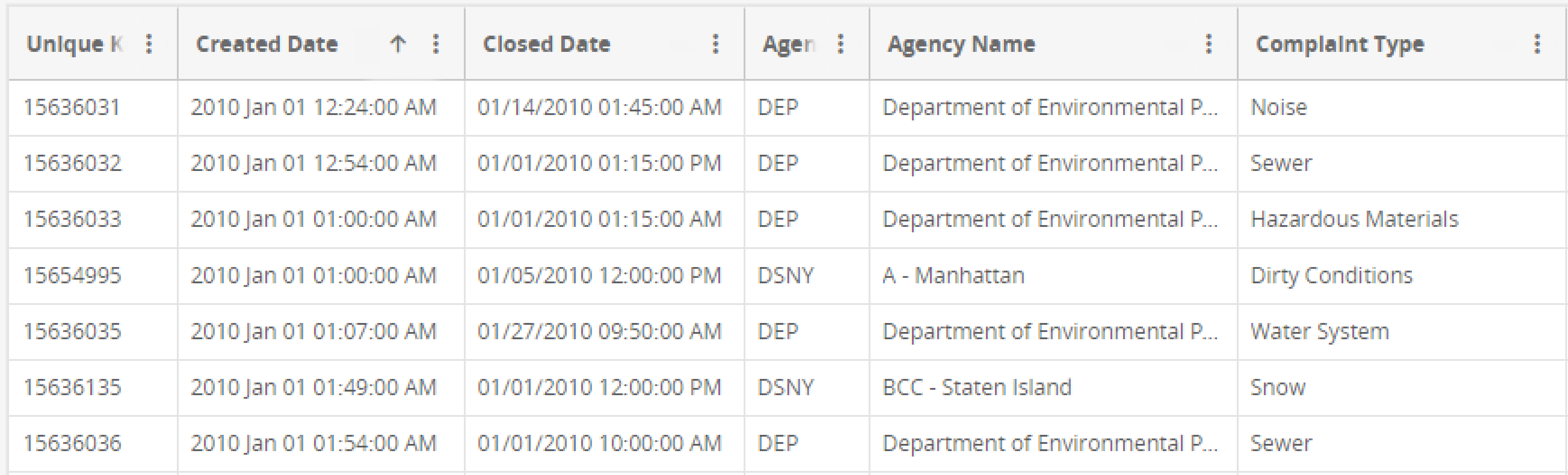
→ Variable importance: potential improvement

Technical Stack: PySpark + Pandas

Data Source

NYC OpenData: 311 Service Requests from 2010 to Present

- https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
- CSV format, 9.36 M rows and 41 columns
  - Unique Key, Created Date, Closed Date, Agency, Complaint Type, Borough …

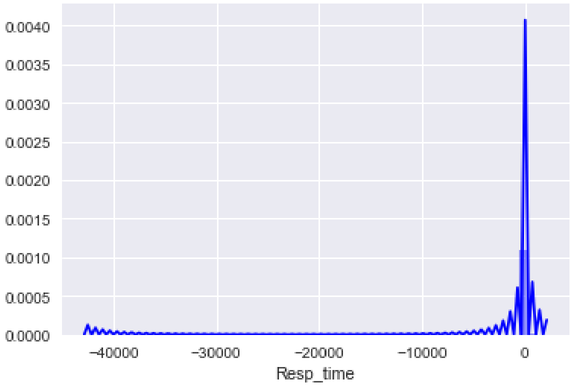


Unique K	Created Date ↑	Closed Date	Agen	Agency Name	Complaint Type
15636031	2010 Jan 01 12:24:00 AM	01/14/2010 01:45:00 AM	DEP	Department of Environmental P...	Noise
15636032	2010 Jan 01 12:54:00 AM	01/01/2010 01:15:00 PM	DEP	Department of Environmental P...	Sewer
15636033	2010 Jan 01 01:00:00 AM	01/01/2010 01:15:00 AM	DEP	Department of Environmental P...	Hazardous Materials
15654995	2010 Jan 01 01:00:00 AM	01/05/2010 12:00:00 PM	DSNY	A - Manhattan	Dirty Conditions
15636035	2010 Jan 01 01:07:00 AM	01/27/2010 09:50:00 AM	DEP	Department of Environmental P...	Water System
15636135	2010 Jan 01 01:49:00 AM	01/01/2010 12:00:00 PM	DSNY	BCC - Staten Island	Snow
15636036	2010 Jan 01 01:54:00 AM	01/01/2010 10:00:00 AM	DEP	Department of Environmental P...	Sewer

Data Wrangling

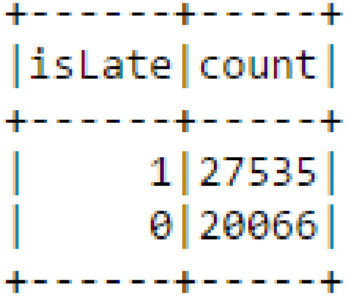
Target Variable: Resp_time (Response Time)

- Resp_time = Closed Date – Created Date [unit: second]
- Issue:
  - Missing values: no Closed Date → eliminate
  - Negative values: Closed Date < Created Date → no explanation, eliminate
  - Hard to read → convert to day



Alternative Target Variable: isLate

- If Response Time > 1 (day), isLate = True, otherwise False
- Quite balanced (show only the sample)

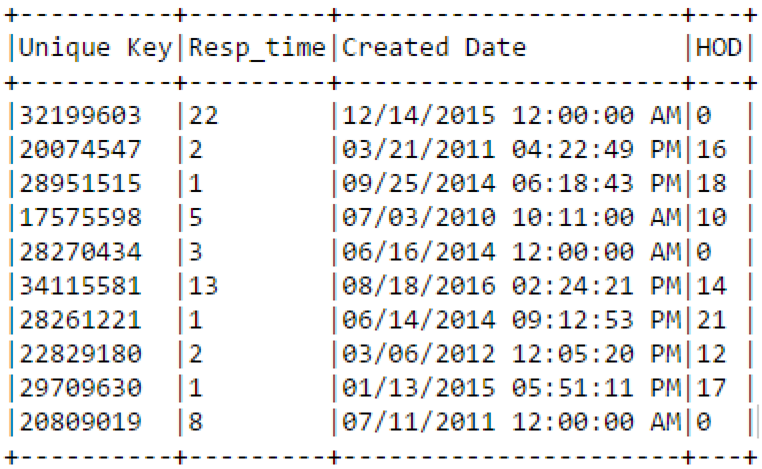


```
+------+-----+
|isLate|count|
+------+-----+
|     1|27535|
|     0|20066|
+------+-----+
```

Data Wrangling

Predictor: HOD (hour of day)

- The time of call matters!
- HOD = hour of Created Date
    - Properly handled the 24-hr format (0-23)



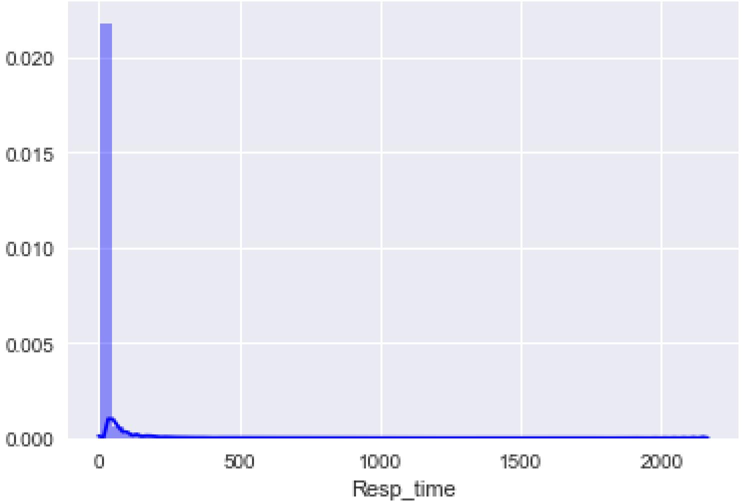
```
+----------+---------+----------------------+---+
|Unique Key|Resp_time|Created Date          |HOD|
+----------+---------+----------------------+---+
|32199603  |22       |12/14/2015 12:00:00 AM|0  |
|20074547  |2        |03/21/2011 04:22:49 PM|16 |
|28951515  |1        |09/25/2014 06:18:43 PM|18 |
|17575598  |5        |07/03/2010 10:11:00 AM|10 |
|28270434  |3        |06/16/2014 12:00:00 AM|0  |
|34115581  |13       |08/18/2016 02:24:21 PM|14 |
|28261221  |1        |06/14/2014 09:12:53 PM|21 |
|22829180  |2        |03/06/2012 12:05:20 PM|12 |
|29709630  |1        |01/13/2015 05:51:11 PM|17 |
|20809019  |8        |07/11/2011 12:00:00 AM|0  |
+----------+---------+----------------------+---+
```

Choice of Variables

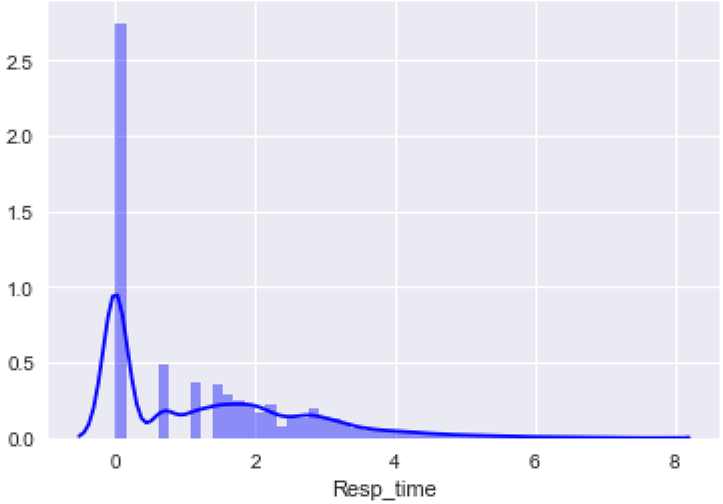
- Eliminate variables having very little information: e.g. Taxi Company
- Choose Borough as the representative variable for geographic information
    - Drop coordinates and other similar variables
- Variables have Unspecified category dominated (95%) → drop

Exploration

Distribution of Resp_time



Distribution of log_Resp_time



Use log_Resp_time instead as the target variable for regression model

Modeling - Regression

Attempt to build a regression model to predict log-response time.

Model tested:

- Linear regression
- Generalized linear regression (Poisson with log-link)
- Random forest regressor
- Gradient boosting tree regressor

Metric: RMSE

Best model: Random forest (100 trees, max depth 10, with max 250 bins)

- Test set prediction RMSE (log-scale): 0.84
- RMSE is too large → Far from satisfaction!

Modeling - Classification

Attempt to build a regression model to predict whether a case will take more than one day to close (isLate).

Model tested:

- Logistic regression
- Naïve Bayes classifier
- Random forest classifier
- Gradient boosting tree classifier

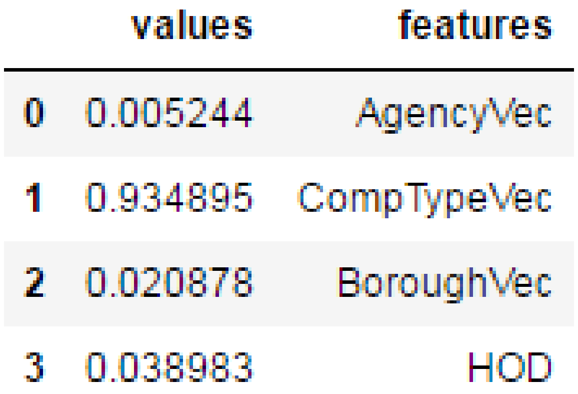
Metric: AUC

Best model: Gradient boosting tree classifier (20 max iteration, max depth 5, with max 250 bins)

- Test set prediction AUC: 0.94

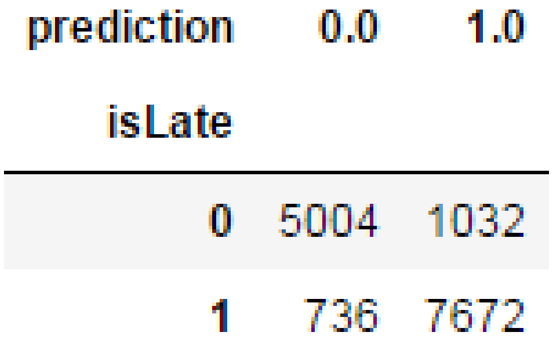
Results

Variable Importance



	values	features
0	0.005244	AgencyVec
1	0.934895	CompTypeVec
2	0.020878	BoroughVec
3	0.038983	HOD

Confusion Matrix



prediction	0.0	1.0
isLate		
0	5004	1032
1	736	7672

Accuracy: 88.3%

Conclusion

Using just a few variables, we are able to predict whether a complaint case may be resolved within one day or tends to be delayed on response with 88% accuracy.

Predicting the actual response time remains a challenge.

The complaint type is identified to be the most important variable, followed by the filing time (although much less impact comparing to complaint type), in determining the response time.