© 2018 SPLUNK INC.

Splunk-ing Crime:

Predicting London Crime Rates using the ML Toolkit

Paul McDonough & Shashank Raina @ **NCC Group**

September 2018 | Version 1.0

© 2018 SPLUNK INC.

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements we may make. In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionality described or to include any such feature or functionality in a future release.

Splunk, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2018 Splunk Inc. All rights reserved.

splunk> .conf18

© 2018 SPLUNK INC.

Your Hosts

PAUL MCDONOUGH

Head of Security & Data Analytics
@ NCC Group

SHASHANK RAINA

Security & Data Analytics Senior Consultant
@ NCC Group

© 2018 SPLUNK INC.

Who are we?

nccgroup

NCC Group is a security consultancy and advisory business helping to solve complex security challenges day in and day out.

Our Ninja's are based worldwide and passionate about making the internet safer and revolutionizing the way in which organisations think about cyber security

United Kingdom
- Manchester (HQ)
- London
- Cambridge
- Edinburgh
- Glasgow
- Leatherhead
- Leeds
- Milton Keynes
- Reading

Europe
- Denmark
- Germany
- Lithuania
- Spain
- Switzerland
- Netherlands (Amsterdam)
- Netherlands (Delft)
- Netherlands (The Hague)

United States & Canada
- Atlanta, GA
- Austin, TX
- Boston, MA
- Campbell, CA
- Chicago, IL
- New York, NY
- San Francisco, CA
- Seattle, WA
- Sunnyvale, CA
- Toronto, ON
- Kitchener, ON

Asia Pacific
- Sydney
- Singapore
- UAE (Dubai)

© 2018 SPLUNK INC.

How can Splunk help with ‘real world’ security problems?

Using Splunk and the ML toolkit to analyze crime pattern behavior

nccgroup

splunk> .conf18

© 2018 SPLUNK INC.

Yes it can!

Splunk: solving real world problems one SPL at a time… ROKT

- Rokt is a martech company connecting brands and e-commerce sites with customers. Rokt helps its clients reach customers at the exact moment they're most receptive – the moment of purchase – to introduce them to something new and relevant. As the business began to grow significantly, Rokt saw an increasing need to centrally manage its logs to increase efficiency and productivity.

Business Impact

- Enhanced business efficiency due to realtime, proactive log management
- Improved productivity with full visibility into IT environment, user behavior and DevOps testing
- New business possibilities created by predictive analytics

Preventing fraud and grasping future opportunities

According to Vermeulen, another game-changing feature of Splunk Enterprise is predictive analytics. "We can generate statistical reports and dashboards at our fingertips and easily spot anomalous events in data," says Vermeulen. "We also predict future activity from the patterns and correlations in the events happening on partners' websites, while sending alerts based on thresholds and enabling 'what-if' scenario analysis."

"Not only has Splunk consolidated all of our log data and pushed it to one central, accessible and easy-to-use interface, but it has also opened the door to new possibilities helping us bring the business to new heights."

Andy Vermeulen, Lead Engineer (DevOps)
Rokt

© 2018 SPLUNK INC.

Why did we do this?

How Splunk and ML Toolkit help us get more from data

- Using Splunk to correlate various datasets to present a holistic image is effortless.
  - We used a lot of publicly available datasets and put them all in Splunk so that we can see the relationship between them easily and in one place.
- Creating Performing Predictive models within Splunk is really straightforward.
  - Using ML Toolkit we can use predictive data models to get a sneak peak into the future with a certain accuracy and be better prepared for any potential issues.

© 2018 SPLUNK INC.

How did we do it….

Let's peak under the hood

Building the Dataset

- We have used publicly available data for Greater London.
  - Data Includes Crime per LSOA, Census for different age groups, School Absences, Child Poverty & Income.
    - We have used data from 2008 to 2011 to build the dataset and then predict for 2012.

Pre-Processing the Data

- To prevent the issues in Model due to large data variations.
  - **kmeans** command to cluster on basis of numerical values.
  - **analyzefields** command to determine the ability for each of those fields to predict the value of a particular field.
  - **anomalousvalue** to compute an anomaly score for each field of each event, relative to the values of this field across other events.
- We then discard the data with which we find a big variance.

© 2018 SPLUNK INC.

How did we do it….(contd.)

Let's peak under the hood

Create the Predictive Model

- We used "Predict Numeric Fields" dashboard in ML Toolkit App.
  - By adding a Pre-Processing step we selected 30 fields from 45.
  - We created 4 models with different algorithms.
  - Then on basis of highest R Squared and lowest mean error values, we chose the model to be used.

Predict Future Values

- Use the model created to predict for future years.
  - The model was created using data from 2008-2011.
  - We will use this model to predict the values for 2012.
  - The model can be improved by putting more data to it.
  - And can also be used for any future predictions.

© 2018 SPLUNK INC.

... And what else?

How to use ML Toolkit to create a Predictive Model

Pre-Processing

- Pre-processing steps transform your machine data into fields ready for modelling or visualization.
- Currently we have 5 options: Field Selector, kernel PCA, PCA, Standard Scaler, TFIDF.
- We have used Field Selector for our experiment as it gives us a good idea of what fields will create a better model.

Algorithms

- For predicting numeric fields we have currently 7 algorithms.
- We have used 4 for our experiment: Linear Regression, Decision Tree Regressor, Lasso & Ridge.
- Overall the toolkit includes over 30 common algorithms and gives you access to more than 300 popular open-source algorithms through the Python for Scientific Computing library.

Predictive Model

- After preprocessing and applying an algorithm a model is created with some attributes.
- The most important is “R Squared”, which is a statistical measure of how close the data is to the fitted regression line.
- Choose the model with highest R Squared but also minimum RMSE (Root Mean Square Error).

© 2018 SPLUNK INC.

What data sources did we use?

Where is this wonderful data from?

London Crime Data

https://data.police.uk/

Office for National Statistics

Census Data:

https://www.ons.gov.uk

LONDON DATASTORE

Various London Datasets

https://data.london.gov.uk

Trust for London

Tackling poverty and inequality

London Poverty Data

https://www.trustforlondon.org.uk/data/child-poverty-borough/

© 2018 SPLUNK INC.

What did we find?

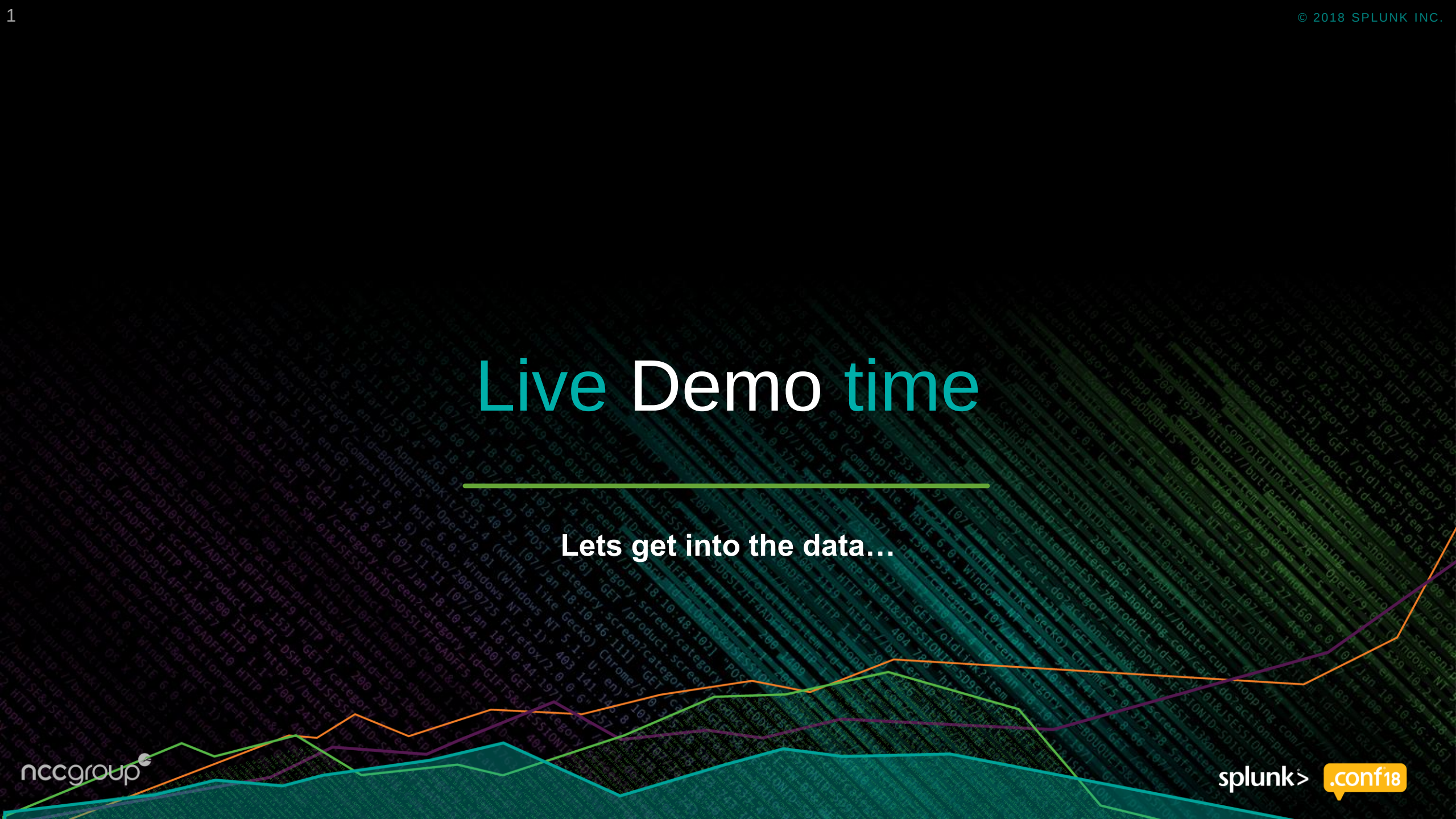
???

- **Splunk & Open Source Data**
  - Creating Performing predictive models with a basic dataset proved remarkably straightforward and didn't require a huge amount of work in Splunk.
  - Splunk provides us with various tools to clean our data which is really necessary for better models.
- **Splunk ML Toolkit**
  - Even without any data science background we can use the ML Toolkit to create predictive data models.
  - By providing analysis of the models, the toolkit helps us to choose the one which is best for our data.
- **London Crime Prediction**
  - As expected, past year offences, total people at all ages, size affect the no. of crimes.
  - But also included are no. of people over 65 and between 16-29 and overall school absence.
  - We don't have control over the predictions. If something significantly changes in the police strategy, law, areas, etc., the model could be really wrong.

© 2018 SPLUNK INC.

Live Demo time

Lets get into the data…



© 2018 SPLUNK INC.

Crime Prediction Model we Created

Crime Prediction Model Per LSOA using Decision Tree Regressor Algorithm

Edit | Export | ...

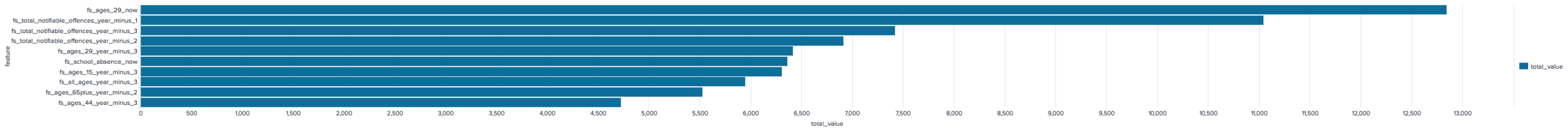
R Square

0.99

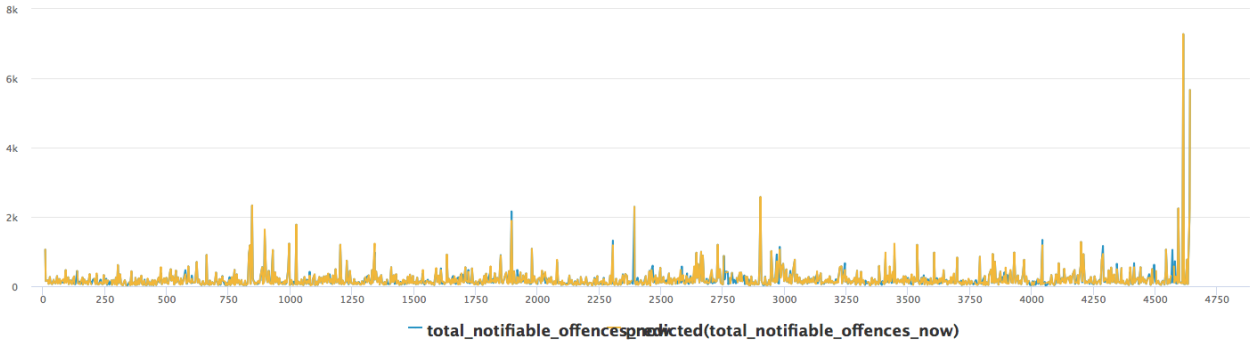
Root Mean Squared Error (RMSE)

26.53

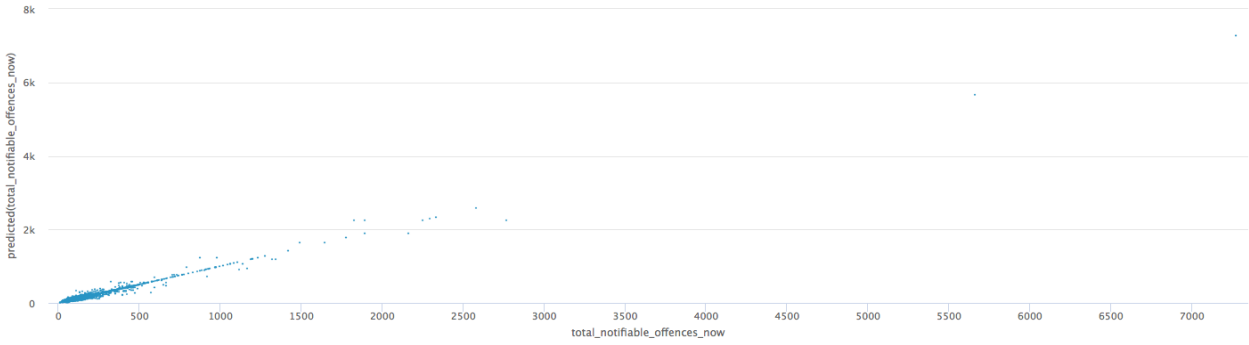
Top 10 Coefficients



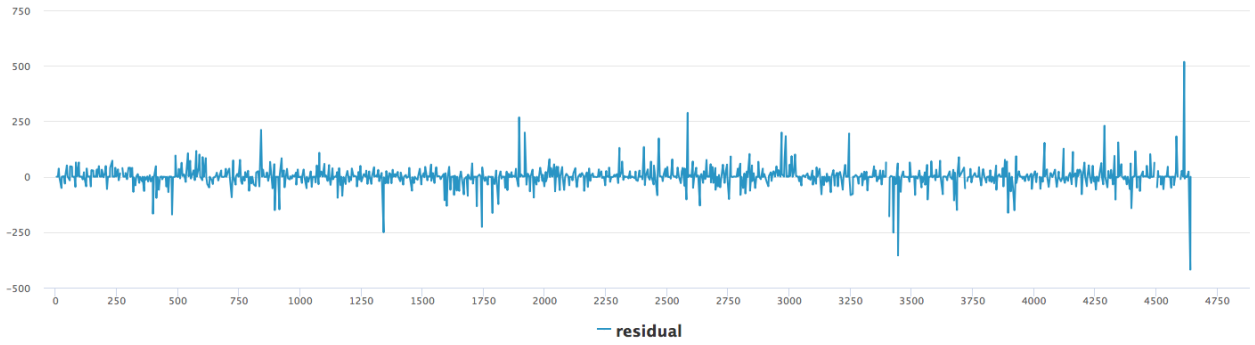
Actual vs. Predicted Line Chart



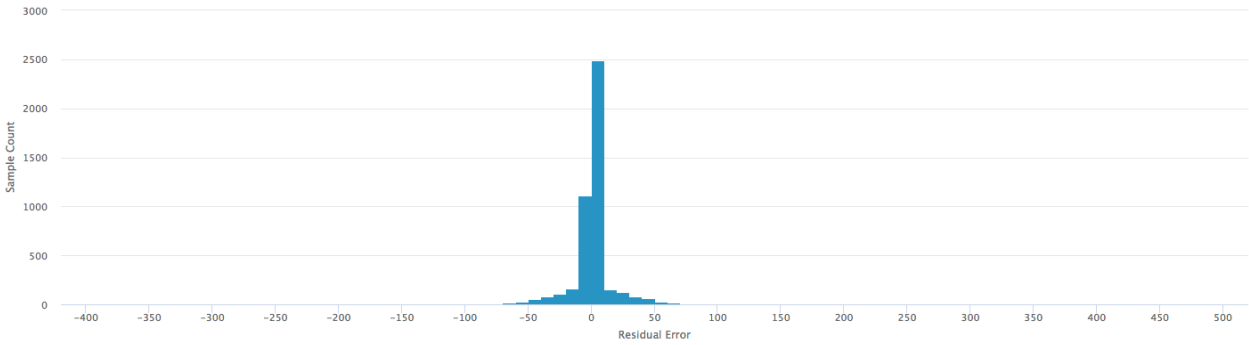
Actual vs. Predicted Scatter Chart



Residuals Line Chart



Residuals Histogram



© 2018 SPLUNK INC.

Crime Prediction for Next Year using the Model

Crime Prediction For Next Year using Decision Tree Regressor Algorithm

Edit | Export | ...

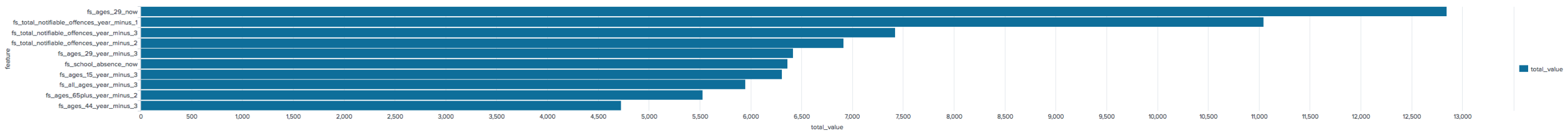
R Square

0.93

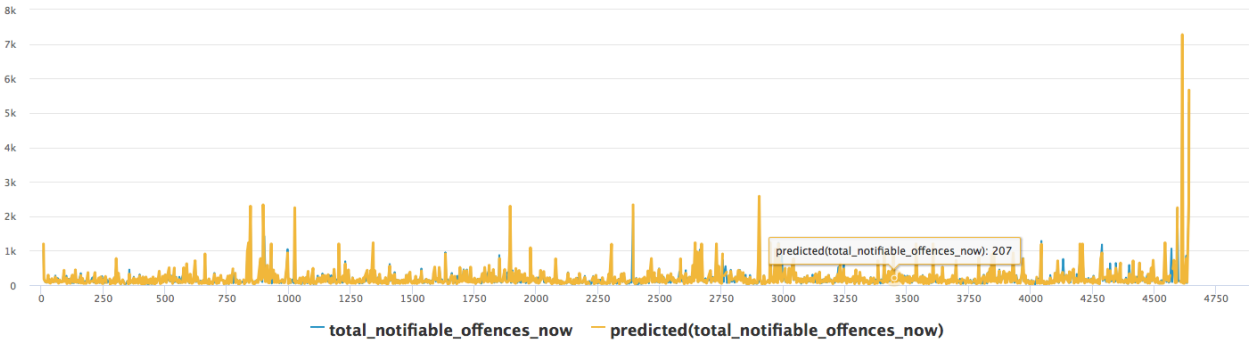
Root Mean Squared Error (RMSE)

53.02

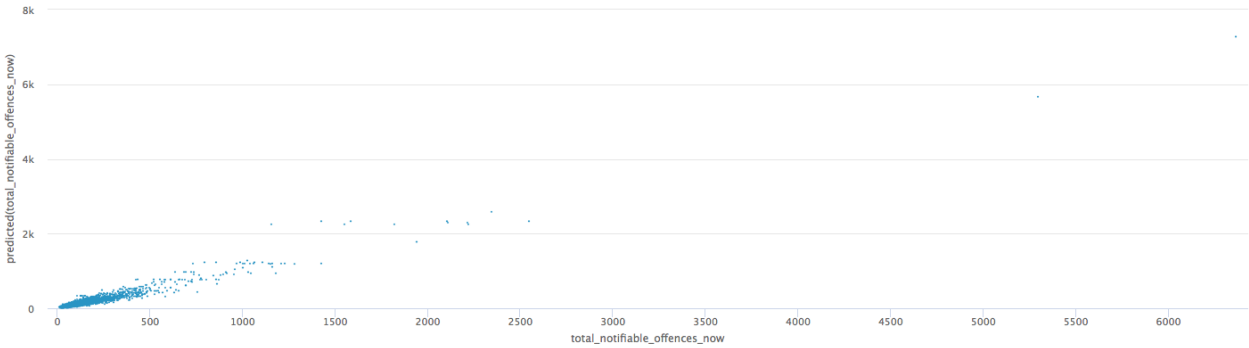
Top 10 Coefficients



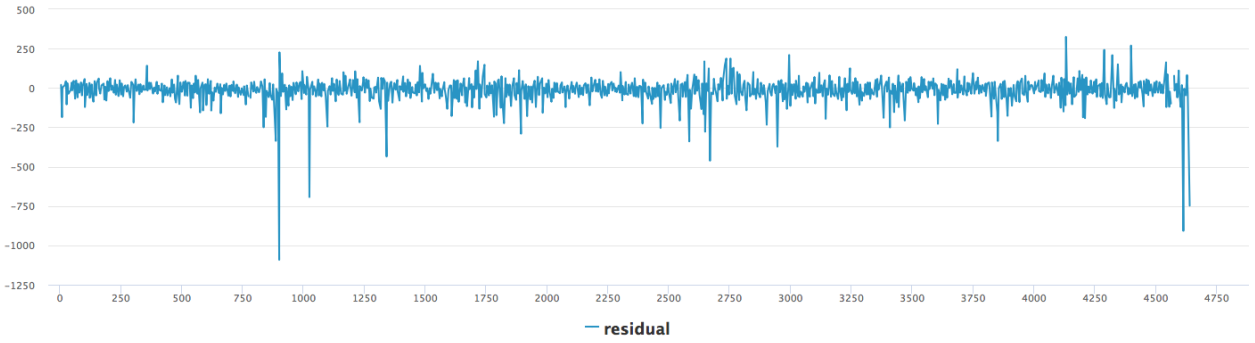
Actual vs. Predicted Line Chart



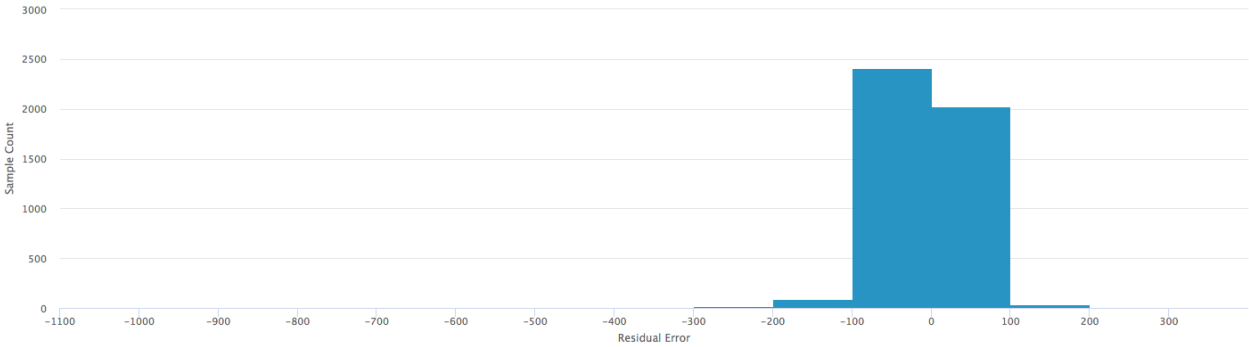
Actual vs. Predicted Scatter Chart



Residuals Line Chart



Residuals Histogram



© 2018 SPLUNK INC.

Future Data source opportunities

What else can we leverage to enhance this capability?

https://openweathermap.org/api

Premier League results

Eventbrite Major Events

Transport for London

And a little of me too!

© 2018 SPLUNK INC.

Key Takeaways

Minority Report time?

1. Splunk **CAN** solve real world problems and show valuable insight into **ANY** data you throw at it…

2. **Machine Learning** isn't just a **buzzword**… it has real world applications when used responsibly

3. Full **technical breakdown** on the demo environment & data **available** on the NCC Group **Blog / Medium** (link tbc)

4. This app is available in the GitHub: https://github.com/nccgroup/Splunking-Crime

© 2018 SPLUNK INC.

Thank You

Don't forget to rate this session in the .conf18 mobile app

nccgroup