From Data Science to Data Stories

Katya Vladislavleva, PhD, PDEng
CEO DataStories (Evolved Analytics Europe)
OUR TECHNOLOGY IS SHAPED BY THE REAL WORLD
Different industries are solving the same analytics problems

- Energy
- Advanced manufacturing
- Materials
- Finance
- (Digital) Health
- Consumer products
- Business operations
If you can measure it, you can understand it.

If you can understand it, you can alter it.

Katherine Neville
OF THE BUDGET GOES TO DATA COLLECTION
95% of the data is not utilized (not monetized).

Source: IDC 2014
More data does not always imply more information
The real goal is understanding and control.
The Age of the Customer is here

“...by 2016, 89% of companies expect to compete mostly on the basis of customer experience.”

“CIOs attach more importance to developing consistent and relevant multi-channel experiences”

“...densely collaborative space between the CIO’s staff and the CMO’s staff.”

“...Nearly half of respondents say their CEOs personally sponsor digital initiatives.”
Good Advice from Forrester

1. Attain Maximum Customer Intelligence with Data Analytics
2. Create customer experience excellence
3. Be there, be digital
4. Embrace the mobile

MISMATCH BETWEEN THE VOLUME OF DATA AND OUR CAPACITY TO ANALYSE IT GROWS ALARMINGLY FAST

By 2018, the US only will face a shortage of up to 190,000 data scientists as well as 1.5 Million managers and analysts with enough proficiency in statistics to use big data effectively.

-McKinsey Global Institute 2013
Define KPIs
Collect Data
Organize & Process Data
Interpret Data
Act on Findings

Competitive Advantage

Profit

Time

Analytics

Analytics

Profit

-$$

+$$

0
Change the Mind set
Change the Mind set

Work in Process costs are reduced by 75%
Change the Mind set

Hmm...
If this is possible,
What if we optimize our products for profitability?

Forward-thinking Leader
There are different kinds of big data

- Big Data
- Medium & Small Data
  - Big Hype
  - Big Mess
The number of options for preparing end users to discover all forms of data grows.

Self-service data preparation tools are exploding in popularity. This is in part due to the shift toward business-user-generated data discovery tools such as Tableau that reduce time to analyze data. Business users also want to be able to reduce the time and complexity of preparing data for analysis, something that is especially important in the world of big data when dealing with a variety of data types and formats. We’ve seen a host of innovation in this space from companies focused on end user data preparation for Big Data such as Alteryx, Trifacta, Paxata and Lavastorm while even seeing long established ETL leaders such as Informatica with their Rev product make heavy investments here.

Additional Reading:
Alteryx, Trifacta, Paxata, Lavastorm, Informatica
Please choose a datasource from the list on the right or type a name in the box above. If you do not have any data sources please upload one using the link below.

**data_extended.csv**

Or upload a new one

If you do not have a datasource, please follow the link to upload a new one.
Predicting Propylene Output (extended dataset)

Nunc pulvinar varius est, quis viverra tortor hendrerit et. Suspendisse facilisis accumsan metus a tempus. Vestibulum eleifend vel nisi eu pulvinar. Vivamus tincidunt vehicula tortor, ac malesuada augue vehicula egestas.
Choose columns we may use and Select your Key Performance Metric

We will analyse your data and build predictive models for the Key Performance Indicator (KPI) you select in the table below. Please, make sure that all columns that we may use as potential predictors are selected in the table below. If there are columns that you do not want to see in the final models (e.g. if they are difficult to measure or control), please, exclude them from the list of options. We will consider all the columns allowed by you and distill a minimal list of necessary and sufficient columns which impact your KPI.

Select your KPI by clicking a KPI button corresponding to the metric of interest. At this point only one KPI at a time is allowed. Let us know at beta@datastories.com if modeling multiple KPIs using the same list of metrics is important for your application. We will make it happen!

data_extended.csv

Showing 1 to 25 of 73 data columns
Give us some context

As much as you can. Context-free solutions lead to context-free results, and we want to make sure your relationship with us is an investment rather than a cost!

What is your application?

What is your critical business objective related to this data? Or not related at all?
Now sit back, relax and give us some time to create your DataStory!

You can WATCH OUR PROGRESS HERE. Our algorithms are very computationally intensive. Finishing all steps of the analysis would take us a minimum of 10 minutes.

You can earn many karma points from us if you send an email with suggestions on how your experience could be made smoother. Please, write us to beta@datastories.com. Katya, Robbe, Sean, Sasha are all checking this email and will respond asap.

You will be redirected to the OVERVIEW PAGE in 2 secs
Quick Summary of your Data

Here is what we got from you

You uploaded `data_extended.csv` file with a filesize of 1.68 Mb on 20.06.2016

<table>
<thead>
<tr>
<th>ROW COUNT</th>
<th>COLUMN COUNT</th>
<th>MOSTLY NUMERIC COLUMNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2199</td>
<td>69</td>
<td>100%</td>
</tr>
</tbody>
</table>

SELECTED KEY PERFORMANCE INDICATOR

Propylene

COLUMNS CONTAINING NUMBERS

69

COLUMNS CONTAINING TEXT

0
Very Simple Relationships
EXPLORE SIMPLE RELATIONSHIPS BASED ON CORRELATION AND MUTUAL INFORMATION

To drill down into how your KPI is connected to the metrics we first checked how all metrics are connected to each other. We found several tightly connected groups of your metrics. We thought you might want to know about inter-relationships.

Below you can play with the first two sliders to see which columns would be connected to each other in terms of correlation or mutual information. If you change the thresholds, how, when the first slider below is set to 20, we draw a line connecting column if their mutual information is greater than 20%. The more you move sliders to the right, the stricter your connection requirements would be, and only super-strong pair-wise relationships will be shown (if any).

Play with a mutual information threshold
Value = 90

Play with a correlation threshold
Predictive Models
A SUMMARY OF PREDICTIVE MODELING RUNS

We had to create and challenge 25579 predictive models to deeply learn which metrics are necessary and sufficient to predict your KPI. A half of the computational effort was spent on meticulous cross-validations to make sure we avoid over-fitting and maximizing the predictive power of models given your data. At the end we have build a final ensemble of 100 models with a minimal number of metrics, which you can use to run interactive "what-if" scenarios.

The final ensemble has the following characteristics:

- Average cross-validation correlation accuracy: 96.8%
- Number of metrics: 3
- Number of metrics we started with: 68
Predictive Models

A SUMMARY OF PREDICTIVE MODELING RUNS
What-If’s
PLAY DIFFERENT WHAT-IF SCENARIOS BELOW TO SEE WHAT HAPPENS TO THE KPI, HOVER OVER THE GRAPHS FOR MORE INFO

Propylene: 0.55

Delta_temperatureTray2_temperatureTray1: 3.10
Delta_temperatureTray4_feedTemperature: -1.50
Delta_temperatureTray7_temperatureTray1: 33.50

deltaPressure | OverheadPressure | OverheadTemperature | refluxFlow | feedTemperature | feedFlow1 | feedFlow2 | temperatureTray1 | temperatureTray2 | temperatureTray3 | temperatureTray4
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
4.04 | 382.0 | -150.0 | 135.0 | 44.6 | 269.0 | 1730.0 | 39.1 | 42.0 | 42.0 | 43.1
4.05 | 382.0 | -150.0 | 135.0 | 44.4 | 264.0 | 253.0 | 38.4 | 41.4 | 41.5 | 42.9
4.05 | 382.0 | -150.0 | 135.0 | 43.2 | 264.0 | 180.0 | 36.8 | 39.6 | 39.8 | 41.2
4.06 | 382.0 | -150.0 | 135.0 | 44.0 | 264.0 | 440.0 | 39.2 | 42.1 | 42.0 | 43.1
4.06 | 382.0 | -150.0 | 135.0 | 42.5 | 263.0 | 271.0 | 36.6 | 39.7 | 39.8 | 40.9
4.06 | 382.0 | -150.0 | 135.0 | 44.0 | 269.0 | 554.0 | 38.4 | 41.3 | 41.3 | 42.5
4.06 | 382.0 | -150.0 | 135.0 | 44.3 | 267.0 | 1040.0 | 38.4 | 41.4 | 41.5 | 42.8
4.06 | 382.0 | -150.0 | 135.0 | 44.3 | 267.0 | 438.0 | 38.7 | 41.6 | 41.7 | 42.8

Showing 1 to 9 of 2,199 entries
What-If's

Play different what-if scenarios below to see what happens to the KPI. Hover over the graphs for more info.

**Propylene: 1.53**

Delta\_temperature\_Tray2\_temperature\_Tray1: 1.88

Delta\_temperature\_Tray4\_feed\_Temperature: -0.10

Delta\_temperature\_Tray7\_temperature\_Tray1: 34.68

<table>
<thead>
<tr>
<th>deltaPressure</th>
<th>OverheadPressure</th>
<th>OverheadTemperature</th>
<th>refluxFlow</th>
<th>feedTemperature</th>
<th>feedFlow1</th>
<th>feedFlow2</th>
<th>temperatureTray1</th>
<th>temperatureTray2</th>
<th>temperatureTray3</th>
<th>temperatureTray4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.04</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.6</td>
<td>269.0</td>
<td>173.0</td>
<td>39.1</td>
<td>42.0</td>
<td>42.0</td>
<td>43.1</td>
</tr>
<tr>
<td>4.05</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.4</td>
<td>264.0</td>
<td>253.0</td>
<td>38.4</td>
<td>41.4</td>
<td>41.5</td>
<td>42.9</td>
</tr>
<tr>
<td>4.05</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>43.2</td>
<td>264.0</td>
<td>188.0</td>
<td>36.8</td>
<td>39.6</td>
<td>39.8</td>
<td>41.2</td>
</tr>
<tr>
<td>4.06</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.0</td>
<td>264.0</td>
<td>440.0</td>
<td>39.2</td>
<td>42.1</td>
<td>42.0</td>
<td>43.1</td>
</tr>
<tr>
<td>4.06</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>42.5</td>
<td>263.0</td>
<td>271.0</td>
<td>36.6</td>
<td>39.7</td>
<td>39.8</td>
<td>40.9</td>
</tr>
<tr>
<td>4.06</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.0</td>
<td>269.0</td>
<td>554.0</td>
<td>38.4</td>
<td>41.3</td>
<td>41.3</td>
<td>42.5</td>
</tr>
<tr>
<td>4.06</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.3</td>
<td>267.0</td>
<td>1040.0</td>
<td>38.4</td>
<td>41.4</td>
<td>41.5</td>
<td>42.8</td>
</tr>
<tr>
<td>4.06</td>
<td>382.0</td>
<td>-152.0</td>
<td>135.0</td>
<td>44.3</td>
<td>267.0</td>
<td>438.0</td>
<td>38.7</td>
<td>41.6</td>
<td>41.7</td>
<td>42.8</td>
</tr>
</tbody>
</table>

Showing 1 to 9 of 2,199 entries
Conclusions of the DataStory: Predicting Propylene Output (extended dataset)

We analyzed Predicting Propylene Output (extended dataset) to assess what drives your key performance metric Propylene using 68 columns you provided. We explored the data health of your data and rated it at 60 in general. Your data had 2199 rows, and the KPI has 158 unique values.

Your data set had none missing cells. With respect to predicting the Propylene the health of your data is 58. At this stage DataStories focus on finding reliable relationships between numeric metrics and your KPI. So, we had to look at 68 metrics remaining after eliminating 0 We first looked at how your metrics impact the KPI individually. For this we performed a standard correlation analysis and a more involved analysis of the mutual information content between Propylene and all other inputs individually. Because the data you provided only had 68 columns on top of the KPI we also computed all individual pairwise relationships (correlation and mutual information) among the metrics to see how things are connected to each other. Based on initial results we could conclude that out of 68, 3 could be removed from the consideration whatever, because they do not have even slight independent relationships to your Propylene.

From this preliminary analysis we could conclude that only 1380 inputs are individually related to your KPI, but many of them are correlated to each other. This means, that you could further improve the model if you remove some of them. We did this and we were surprised by the performance. After deeply learning your prediction problem and having created and challenged 25579 models, we discovered that 3 are sufficient to predict your Propylene at 99% correlation. These driver metrics have various influence on the KPI and have to be used together to make robust predictions. The drivers are Delta_temperatureTray2, temperatureTray2 (Importance: 60%), Delta_temperatureTray3, feedTemperature (Importance: 60%), Delta_temperatureTray4, feedTemperature (Importance: 48%), Delta_temperatureTray7, temperatureTray1 (Importance: 48%), altogether their importances sum up to 100%. You can play with how they impact Propylene in the What-if scenario tool (here).

If by exploring the drivers you realized that some of them are very difficult to measure or control, or might be coupled with your performance, try to re-run the DataStory while eliminating them from the list of candidate metrics within the DataStory setup. Know when we have the models we can identify outliers, or optimize the models to find optimal settings to achieve desired Propylene levels. This is a premium feature, please, contact us to discuss this.

We are working very hard to add model evaluation functionality and model export functionality and for now you can upload the data with empty KPI values, and we will fill them in with predictions.

Let us know how you liked it!

DataStories
Challenges

- Predictive Models
  - Data-Driven Innovation guiding Sustainability
- Interactive Dashboard

P&G

Datastories
### Process Data: 154,260 energy observations

<table>
<thead>
<tr>
<th></th>
<th>TOTAL NUMBER OF TAGS</th>
<th>TAGS WITH AT LEAST 50% DATA</th>
<th>TOTAL NUMBER OF ENERGY OUTPUTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT TAGS</td>
<td>1,239</td>
<td>1,183</td>
<td>12</td>
</tr>
<tr>
<td>DISCRETE TAGS</td>
<td>292</td>
<td>609</td>
<td>574</td>
</tr>
<tr>
<td>CONTINUOUS TAGS</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Our Approach to the Hackathon Challenge

1. Collect logged measurements from all tags
2. Process, organize, and aggregate the data
3. Run predictive modeling, find energy consumption drivers and predict energy consumption
4. Deploy predictions in a dashboard with interactive what-if scenarios
Our Hackathon Outcomes

1. Built Non-linear Predictive Models for Power Consumption with local error bounds

2. Which lead to a Dashboard with focus on the variables that currently matter most

3. That also integrates an innovative approach to run What-If Scenarios starting from the current process state

4. And allows you to maximize the Sustainability of your process at all times.
Data Pre-Processing & Analysis

✓ Integrated and aligned all available tags with energy data
✓ Generated monthly and quarterly datasets with 5 min averages
✓ Looked at Data Health, Data distributions, Linear Correlations, Mutual Information Content and Variable Connections & Grouping
Interactive Explorer of Tag Relationships
Our Modeling Process

1. Organize all your process data into one big table
2. Define your KPIs
3. Apply advanced machine learning
4. $$$ Identify driving metrics & Deploy Predictive Models $$$
Build Compact Non-linear Models per regime using extensive process of variable competition and elimination.
Result is compact 5-variable model-ensembles with error limits for each regime.
Dashboard with a focus on Power Consumption
Models are Robust by Design

✓ 600,000 models created to produce a final ensemble for one region
✓ 2 regions of electricity consumption lead to 1,200,000 models
✓ 50% of the modeling effort is spent on cross-validation and making sure the models are predictive and not over-fitting
✓ From 1,183 potential inputs only five (5) metrics per region are necessary and sufficient
✓ Global R2 0.97; per regime R2 is 0.68-0.7
✓ Final ensembles consist of 100 models each and also provide confidence limits
Time It took Us

Total Time Spent: 71 h

- Data Download
- Data Preparation & Aggregation
- Predictive Modeling (not counting DataStories’ RoboElves)
- Matlab, Python, JS, HTML5
- Data Visualization & Connected Dashboard
- Presentation Prep
- Meetings & Driving
Benefits of using Matlab

✓ Super fast implementation
✓ Reliable deployment + flexibility (Matlab package, Stand-alone, Cloud)
✓ Code protection
✓ Matlab users can integrate it easily in their Matlab-based routines
✓ Plug&Play Hadoop integration
Business Outcomes

✓ Project of high business value
✓ Perfect product validation

Special Thanks to Brussels Data Science Community for organizing the hackathon