Data Science SF

Data science for service change

Presented by Tania Jogesh
Data Scientist
City and County of San Francisco
@datasf | datasf.org
Agenda

- Welcome from the CDO
- What is data science?
- What is DataScienceSF?
- Quiz!
- What can data science do for you?
- Activity: Start your application!
- Next steps
What is Data Science?
Using models built on existing data to generate insights and predictions
Data science is a component of analytics

- A live dashboard built on clean automated datasets can get you a long ways!
- But in some cases, it makes sense to use a more complicated model to understand a problem or to make a prediction.
Dashboarding is distinct from Data science

Dashboards
- Communicate insights
- Show past trends
- Based on “simple” analysis

≠

Models
- Generate insights
- Continuous learning
- Predict & prescribe
- Complex analysis
Data science models are everywhere!

We engage with data science models on a regular basis:

- Amazon recommendations
- Uber/Lyft rides
- Chatbox/Virtual Assistant
...and are expanding to public services
What is DataScienceSF?
What is DataScienceSF?

Data Science
Applying advanced statistical tools to existing data to generate new insights

Service Change
Converting new data insights into (often small) changes to business processes

Smarter Work
More efficient and effective use of staff and resources & better services for residents
<table>
<thead>
<tr>
<th>Approach</th>
<th>Process</th>
<th>Outcome</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance Management</strong></td>
<td>Define, visualize, often using dashboards, and manage to KPIs</td>
<td>Meet goals and KPI targets</td>
<td>SF Scorecard, PublicWorks Stat &amp; Stat starter kit</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
<td>Assess a project, program or policy design or results</td>
<td>Better investment of resources; Better policy decisions</td>
<td>Evaluation of transitional-kindergarten in SF</td>
</tr>
<tr>
<td><strong>Policy Analysis</strong></td>
<td>Define and assess alternatives using a broad range of tools</td>
<td>Report or memo with policy or program recommendations</td>
<td>Shape Up SF Policy Analysis</td>
</tr>
<tr>
<td><strong>Open Data</strong></td>
<td>Publish civic data for use by the City and the public</td>
<td>Easier data sharing and reporting, new tools or services built on data</td>
<td>SFPUC Adopt a Drain</td>
</tr>
<tr>
<td><strong>DataScienceSF</strong></td>
<td>Identify insights using advanced statistics tied to a service change</td>
<td>Smarter work “on the ground” in real time</td>
<td>See rest of deck!</td>
</tr>
</tbody>
</table>
What complements data science?
(and is really good stuff to do)

<table>
<thead>
<tr>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Management</td>
</tr>
<tr>
<td>Evaluation</td>
</tr>
<tr>
<td>Policy Analysis</td>
</tr>
<tr>
<td>Open Data</td>
</tr>
<tr>
<td>DataScienceSF</td>
</tr>
</tbody>
</table>

All of these approaches are important and usually interdependent!
What’s in the DataScienceSF Toolkit?

**Statistical Methods**
- Multilevel modeling
- Survival analysis
- Missing data imputations
- Pattern recognition
- Logistic, multinomial, and multiple linear regression techniques
- Classification and clustering
- Principal component and factor analysis
- Forecasting
- Network analysis

**Tools**
- Time series analysis
- Data mining

**User Experience Research**
- AB testing
- Propensity score matching
- Logistic, multinomial, and multiple linear regression techniques
- Network analysis
## What’s in the DataScienceSF Toolkit?

### Tools

<table>
<thead>
<tr>
<th>Statistical Methods</th>
<th>Tools</th>
<th>User Experience Research</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Languages</strong></td>
<td><strong>Libraries</strong></td>
<td><strong>Data Engineering</strong></td>
</tr>
<tr>
<td>Python</td>
<td>SciPy</td>
<td>Profiling</td>
</tr>
<tr>
<td>R</td>
<td>Pandas</td>
<td>ETL</td>
</tr>
<tr>
<td>SQL</td>
<td>Scikit-learn</td>
<td>Job notices</td>
</tr>
<tr>
<td>Javascript</td>
<td>GPText</td>
<td>APIs</td>
</tr>
<tr>
<td>NodeJS</td>
<td>OpenNLP</td>
<td>Optimized data pipelines</td>
</tr>
<tr>
<td></td>
<td>Mahout</td>
<td>Optimized data storage/access</td>
</tr>
<tr>
<td></td>
<td>+many others</td>
<td></td>
</tr>
</tbody>
</table>
What’s in the DataScienceSF Toolkit?

**Statistical Methods**
- Iterative Prototyping
- Journey mapping
- Ethnographic field research and user observation

**Tools**
- User Experience Research
  - Photo journaling and documenting
  - Service blueprinting
- Ride-alongs
- Process mapping
- Usability testing
What is NOT data science?

✅ This

- Service change
- Small changes
- Use existing data

❌ Not that

- Academic research
- Major overhauls / service disruptions
- Collecting new data (mostly ;)

✓ This

☑️ This

☒ Not that
In addition to the service change, we emphasize equity

- many sophisticated models can have inequitable outcomes
- we prioritize projects that have a positive equity impact
What can data science do for you?
Data science project types

1. Find the needle in the haystack
2. Prioritize your backlog
3. Flag “stuff” early
4. Optimize your resources
5. A/B test something
<table>
<thead>
<tr>
<th></th>
<th>Data science project types</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Find the needle in the haystack</td>
</tr>
<tr>
<td>2</td>
<td>Prioritize your backlog</td>
</tr>
<tr>
<td>3</td>
<td>Flag “stuff” early</td>
</tr>
<tr>
<td>4</td>
<td>Optimize your resources</td>
</tr>
<tr>
<td>5</td>
<td>A/B test something</td>
</tr>
</tbody>
</table>
Project Type: Find the needle in the haystack

What to target?

Service Issue: Difficult to identify targets in a population

Data Science

Data Science Process: Use existing data and predictive modeling to identify targets

Service Change

Service Change: Engage with target subset of population
Increasing opt-ins for PUC’s CleanPowerSF Supergreen program

SuperGreen

100% RENEWABLE

UPGRADE TO SUPERRGREEN
Who should they target to increase supergreen sign-ups?
Used existing data to segment customers and identify groups most likely to opt-in

- Customer locations, energy usage data
- Most likely to opt in were residents in the highlighted neighbourhoods and those who had opted for a PG&E “Time of Use” rate
Insights will help them design materials and A/B test their communications
Data science project types

1. Find the needle in the haystack
2. Prioritize your backlog
3. Flag “stuff” early
4. Optimize your resources
5. A/B test something
Project Type: Prioritize your backlog

What to prioritize?

Data Science

Service Change

Service Issue: Backlog is tackled via first in, first out (FIFO)

Data Science Process: Create a model to categorize and group past and current cases

Service Change: Prioritize cases based on categories in order of risk, need or opportunity
Streamlining property tax appraisals

Sale price

Fair market value
Sale price may not always reflect fair market value...

Ms. Jenkins

Sale to friend = $500K

Sale on open market = $1M
Doing a full appraisal for each sale is time consuming and ASR has heavy caseloads.
We developed models that generate a predicted fair market value.
If the sale price is far away from the predicted price, ASR conducts a full appraisal.
Result = Increased revenue!

⇒ The first run of the models reduced the workload by 166 properties or 10% giving the city immediate access to $239 million in roll value leading to ~$2.8M in revenue

+$2.8\ M\ in\ tax\ revenue
## Data science project types

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Find the needle in the haystack</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Prioritize your backlog</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Flag “stuff” early</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Optimize your resources</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>A/B test something</td>
</tr>
</tbody>
</table>
Project Type: Flag “stuff” early

Service Issue:
Hard to predict future condition which leads to reactive services

Data Science Process:
Use historical and current data to create estimate ranges for potential outcomes

Service Change:
Use estimates to change and tailor intervention points
Improving outcomes for residents in the Workforce program

JobsNOW!

JobsNOW! provides employment and training services to income-qualified San Franciscans and offers wage reimbursement to employers when they hire program participants.
HSA Goal

Help HSA’s Jobs Now program target its efforts and resources toward opportunities that are most likely to lead clients toward career advancement.

Data science can help to understand which industry, employer, occupations resulted in better client outcomes.
Incorporate labor market lens into outreach & strategy efforts

- HSA should incorporate an industry strategy for outreach efforts for future employers

Earnings increased with:

- Employer size
- Industry rep in SF
- Healthcare jobs
- Admin roles
These insights helped HSA redesign their jobsNow program

Untargeted partnerships with employers  \rightarrow  Targeted partnerships that improve potential earnings for clients
Data science project types

1. Find the needle in the haystack
2. Prioritize your backlog
3. Flag “stuff” early
4. Optimize your resources
5. A/B test something
**Project Type:** Optimize your resources

**How to distribute?**

**Data Science**

**Service Change**

**Service Issue:**
Difficult to identify where to place or distribute resources to be most effective

**Data Science Process:**
Use geospatial and/or other data to identify optimal distribution of resources

**Service Change:**
Re-allocates resources to optimal distribution
Where are PW sensors likely to break?
Estimated that 50% of sensors will break after 33 weeks overall

- Steady decrease in working sensors
- Are there specific places that sensors are breaking?
Sensor survival rate varies by neighborhood

- In Pacific Heights, Inner Richmond, and Bayview sensors survive the shortest amount of time.
- Cellular connectivity and vandalism could be reasons why.
Insights provided guidance for procuring new sensors

- Current sensors were breaking frequently and having cellular connectivity issues
- New contracts should include failure reason (vandalism, missing, etc) to track issues for easy reporting
<table>
<thead>
<tr>
<th></th>
<th>Task Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Find the needle in the haystack</td>
</tr>
<tr>
<td>2</td>
<td>Prioritize your backlog</td>
</tr>
<tr>
<td>3</td>
<td>Flag “stuff” early</td>
</tr>
<tr>
<td>4</td>
<td>Optimize your resources</td>
</tr>
<tr>
<td>5</td>
<td>A/B test something</td>
</tr>
</tbody>
</table>
Project Type: A/B test something

Service Issue: Costly outreach methods are not tested before implementation

Data Science Process: Statistical testing on outreach methods to identify which, when, and to whom to send

Service Change: Use statistically validated outreach method

Which form?

Data Science

Format A

Format B

62% respond

78% respond

Service Change

Format B
TTX: Increase response to tax letter

Service Issue
TTX wanted to use behavioral economics and A/B test to increase effectiveness of collection letter for unsecured personal property (a traditionally difficult type to collect on).

Data Science
DataSF helped organize a Behavioral Insights Training (BIT) workshop and provided guidance on A/B test.

Service Change
Use whichever letter gets the best response.

Result
Improved response rate by 17%. TTX continuing to apply BIT principles to other taxpayer communications.
Activity

• Take 5 minutes in your breakout room
  – Introduce yourself
  – Discuss: What are some potential projects you think data science could help with?
Next steps for you
Overview of Project Phases

Cohort 5
Aug 2022 - Sept 2023

2022
Aug-Oct  | Oct 7   | Oct  | Nov 7  | Nov-Dec | Jan-Jul*  | Sept

Application period
Application due!
Project selection
Applicants notified!
Cohort kickoff
Project delivery

Charter sign off, analysis, modelling, & service change

*We work with you to agree project timing

Present findings & lessons learned
Phase: Solicitation

Opportunities to learn more:

• Brown bags (8/12 and 9/23)
• Office hours
• Ad hoc presentations

Dates at datasf.org/science
Phase: Application

Brief online form

- Problem statement (200 words)
- Impact statement (100 words)
- Service change statement
- Data overview
- Due Oct 7

Available at datasf.org/science
Successful applications are made during office hours!

- Brainstorm projects
- Get lingering questions answered
- High correlation between strong applications and prior attendance of office hours
Phase: Selection

Process

- Initial review
  - Criteria assessment
  - Application scoring
- Department follow-ups, as needed
  - Be available for questions

Read about our selection process
Phase: Winners Announced

And gentle off-ramps for the rest...

Some projects may not be appropriate for data science or for our timeline. We will help identify other opportunities that may be a better fit:

- Civic Bridge
- STIR
- DataSF Dashboarding Services
- Controller's Performance Unit
- Data Academy classes (Fall 2022!)
- External Data Science groups or volunteers
- Other technical assistance
Selected? Key things to know

• **Service is free!**

• **Dept provides project champion**
  – Prepare for, on average, **25% time spent on DataScienceSF**
  – Some weeks less, some weeks more
  – Act as our main point of contact through the DataScienceSF process

• **Project logistics**
  – Projects tackled sequentially, not concurrently
  – 1-2 months on average per project
  – **During kick off meeting we will set order (you have a say!)**
Phase: Analysis and service change

Iterative Analysis*

- Implementation Research
- Implementation Pilot
- Model Build Out
- Model Evaluation
- Final Model/Handoff
- Document & Disseminate

*Things can change, project team adapts to meet goals and improve service delivery
Ethics & Algorithm Toolkit

Why is this important?

- Outcomes can have important consequences for people in the city
- Understand biases in the data so we can use data responsibly
- Understand biases in the models so we can use algorithms responsibly

How?

Done jointly with DataSF & Department at start of project

Developed in partnership with:
- DataSF
- Harvard Ash Center
- John Hopkins Center for Government Excellence
- Data Community DC
Phase: Analysis and service change

What DataSF Brings
- Statistical Methods
- Tools
- User Experience Research

What You Bring
- Issue expertise
- A good question + data
- Project champion

Final Product: Algorithm + Tool
- scripted and automated algorithm
- tied to some service change tool (e.g. list, service, alert)
- implemented together

Timeline:
- Aug-Oct
- Oct 7
- Oct
- Nov 7
- Nov-Dec
- Jan-Jul
- Sept
Your next steps:
Visit datasf.org/science:
• Sign up for office hours
• Apply by Oct 7!

Questions?
email/ Teams: tania.jogesh@sfgov.org
Meet the team & acknowledgements

Michelle
Chief Data Officer
...and data mastermind

Tania
Data Scientist
...and bug doctor

Cody
Analytics Strategist
...and jack of all trades

Helen
Analytics Engineer
...and tooling master

Many thanks to New Orleans’ NOLAlytics team, New York City’s MODA, and Harvard’s DataSmart for their resources, tools and templates!
Data, for the love of the City

Thank you! Questions?
@datasf | datasf.org | datasf.org/blog