Mastering the Proof of Concept

Brendan.Herger@capitalone.com

Slides: goo.gl/GCniKf

Who am I?

Brendan Herger

- Machine Learning Engineer @ Capital One
- Founding member of the Center for Machine Learning @ Capital One
- I've lead projects including: Cloud infrastructure, fraud detection, and legal / risk categorization

Agenda

What is a POC?

Before, During & After

Case Study

What is a POC?

- Sample project to learn project domain and show value
- "Build one to throw away"
- Key outcomes: Identify that project is possible, adds value, and lays groundwork for longer project

Before

Set yourself up for success, set expectations

Data

- Identify relevant data & access it
- QA a sample of the data
- Identify data size (single machine, many machine, many clusters)

Customer

- Identify project champion, technical domain expert
- Have dinner & drinks together; these people define your success.

Contributors

- Agree to project goals, methods of communication, and conflict resolution
- Have dinner & drinks together; trust can make or break a project.

During

Inspect and adapt; Demo or die

Data

- Decide on a centralized data store (e.g. local FS, HDFS, SQL)
- Version data sets

Customer

- Frequently get feedback
- Archive additional asks / project ideas

Contributors

- Own specific verticals
- Accept technical debt; this one's to throw away anyways!

After

Set up the next team for success, over emphasize handoff

Data

- Capture data sources, owner point of contacts, potential additional data sources
- Identify which data sets should be put into long term storage
- Remove access / permissions as necessary

Customer

- Communicate all of your successes
- Highlight additional asks you've received over the course of the project
- Ask for honest feedback

Contributor

- Dissect the project. What went well? What could go better next time?

Case Study: Customer Churn

Building a customer churn model to help customer relations and marketing reduce churn

Case Study: Customer Churn

Before

Data

- Create a list of fields we'd like
- Access data, pull samples to cluster, QA data

Customer

- Identify their hypothesis about what can be improved (e.g. 'Our app sucks!')
- Identify what they can control, and the kinds of changes they're willing to make
- Determine who is your champion, and what their goals are for the project

Contributors

- Decide on working in Python on a single, shared development machine
- Build your team!

Case Study: Customer Churn

During

Data

- Write all intermediate and final tables to Hive
- Document all data sets in markdown files in the project repo

Customer

- Demo once a week
- Get feedback (e.g. "it's great to know that our customers churn regularly, but that doesn't help...")

Contributors

- Have on contributor focus on data ETL, another on model building, and another on UI
- Define interfaces ahead of time

Case Study: Customer Churn

After

Data

- Write all Hive tables to CSV, and cache somewhere for future reference
- Revoke all access to Hive cluster

Customer

- Explain how to lower customer churn
- Estimate cost for a full project
- Ask them to meet with your product manager to describe what went well and what didn't go well

Contributors

- Decide to keep Hive for next project, try Scala

Thank you!

Brendan.Herger@capitalone.com

Slides: goo.gl/GCniKf

We're hiring!