A Data Science Platform

Vishal Doshi
vishal@celect.com
@superdosh

Akshay Anand
Sai Kiran Burle
Tony Do
Ying-zong Huang
Peter Lin
Michael Lowney

Ritesh Madan
Balaji Rengarajan
Ben Schoener
Devavrat Shah
Jae Sim
Daniel Xu
Prediction Problem Set

• Retail
  – Predict the demand for new products next quarter based on historical transactions, as well as product attributes
  – Predict the demand of products at stores that never carried them
  – Predict the daily demand of products at a store based on historical transactions

• Maritime
  – Predict likely paths for commercial shipping vessels based on AIS data
  – Predict the likely flag of a ship given its path

• Geopolitics
  – Predict future relations between countries along various axes based on an NLP-generated dataset (e.g. GDELT)
Approach

• High-level data-driven decision making

  – Step 0. Identify the problem
  – Step 1. Identify the available data
  – Step 2. Identify the “prediction” question (to solve)
  – Step 3. Make predictions to support decisions
Status Quo

Varied Data Sources → Manual Data Processing → Data Normalization → Modelling Machine Learning → Predictive Queries → Optimized Decisions
pDB: prediction DataBase

Varied Data Sources

Flexible Data Connectors

prediction DataBase

Predictive Queries

Optimized Decisions
Movie Recommendation System

• Building a Movie Recommendations System
  – Step 0. Recommend movies to users that s/he likes
  – Step 1. MovieLens dataset
  – Step 2. Predict what a user will rate a movie
  – Step 3. Build pipelines in pDB
MovieLens Data

• Basics:
  – 27K movies, 138K users

• Ratings: (userId, movieId, rating)
  – 20M ratings (0.53% density)

• Movies: (movieId, title, genre)
  – 27K movies

• Tags: (userId, movieId, tag)
  – 465K tags (free form text)

<table>
<thead>
<tr>
<th>userId</th>
<th>movieId</th>
<th>rating</th>
<th>timestamp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3.5</td>
<td>1112486027</td>
</tr>
<tr>
<td>1</td>
<td>151</td>
<td>4</td>
<td>1094785734</td>
</tr>
<tr>
<td>91</td>
<td>3066</td>
<td>3</td>
<td>1111558027</td>
</tr>
</tbody>
</table>
MovieLens Data

- Basics:
  - 27K movies, 138K users

<table>
<thead>
<tr>
<th>moviel</th>
<th>title</th>
<th>genres</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toy Story (1995)</td>
<td>Adventure</td>
</tr>
<tr>
<td>2</td>
<td>Jumanji (1995)</td>
<td>Adventure</td>
</tr>
<tr>
<td>3</td>
<td>Grumpier Old Men (1995)</td>
<td>Comedy</td>
</tr>
<tr>
<td>4</td>
<td>Waiting to Exhale (1995)</td>
<td>Comedy</td>
</tr>
<tr>
<td>5</td>
<td>Father of the Bride Part II (1995)</td>
<td>Comedy</td>
</tr>
</tbody>
</table>

- Ratings: (userId, movieId, rating)
  - 20M ratings (0.53% density)

- Movies: (movieId, title, genre)
  - 27K movies

- Tags: (userId, moviel, tag)
  - 465K tags (free form text)
MovieLens Data

• Basics:
  - 27K movies, 138K users

<table>
<thead>
<tr>
<th>movielid</th>
<th>title</th>
<th>genres</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Toy Story (1995)</td>
<td>Adventure Animation Children</td>
</tr>
<tr>
<td>2</td>
<td>Jumanji (1995)</td>
<td>Comedy Fantasy</td>
</tr>
<tr>
<td>3</td>
<td>Grumpier Old Men (1995)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Waiting to Exhale (1995)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Father of the Bride Part II (1995)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>movielid</th>
<th>title</th>
<th>genres</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>208 dark hero</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>353 dark hero</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>521 noir thriller</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>592 dark hero</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>668 bollywood</td>
<td></td>
</tr>
<tr>
<td>65</td>
<td>898 screwball comedy</td>
<td></td>
</tr>
</tbody>
</table>

• Tags: (userId, movielid, tag)
  - 465K tags (free form text)
pDB Model of the World

pDB Language

\((\text{operation, (id1, id2)}) : \text{value}\)

Prediction

\((\text{operation, (id1, id2)}): ?\)
pDB Model of the World

pDB Language

\[(\text{rating}, (\text{user}_1, \text{movie}_n)) : 5\]

Prediction

\[(\text{rating}, (\text{user}_1, \text{movie}_2)) : ?\]
pDB Model of the World

pDB Language

\[(\text{rating}, (\text{user}_1, \text{movie}_n)) : 5\]

Prediction

\[(\text{rating}, (\text{user}_1, \text{movie}_2)) : ?\]

- \[(\text{tag}, (\text{user}_1, \text{movie}_1)) : \text{‘screwball comedy’}\]
- \[(\text{genre}, (\text{movie}_1)) : \text{‘Comedy’}\]
- \[(\text{title}, (\text{movie}_n)) : \text{‘Casablanca’}\]
Crossvalidation

RMSE 0.860 0.825 0.793
pDB Model of the World

• The language can simply express typical data science problems
  
  - Regression
  - Classification
  - Time Series (Interpolation, Forecasting and Multiple)
  - Matrix Completion
  - Tensor Completion
pDB Model of the World

• Prediction problem
  – 3-order tensor completion
  – with component being vector valued

• Non-parametric view
  – $T_{op, id1, id2} = f(x_{op}, y_{id1}, z_{id2})$

• Similarities through latent features
  – id1 vs id2 via $y_{id1}$ vs $z_{id2}$
Solutions using pDB: Retail, Federal
Decision Making in Retail using pDB

**Plan**
- Budget Allocation
  - Am I investing in the right areas?

**Buy**
- Optimized Assortments
  - Will I over-buy or under-buy?

**Allocate**
- Optimized Allocation
  - Is the product in the right stores?

**Sell**
- Engaged Customers
  - Will I maximize full price sell-through?

**Liquidate**
- Optimized Markdown
  - How much margin will we lose?
Decision Making in Retail using pDB

• Use pDB to *stitch* data across people, products, locations, time
Maritime Domain Awareness: Anomaly Detection
Maritime Domain Awareness: Predicted Destination
Maritime Domain Awareness: Anomalous Behavior
Enterprise Grade Architecture

• Micro-services architecture leveraging gRPC and protobuf

• Stateless services with end-to-end fault recovery

• Datastores, models and Spark clusters managed seamlessly

• pDB deployed as Docker containers via Kubernetes
pDB is versatile

• High bandwidth connectors integrate in existing environment
  – You do not need to upload your data

• pDB language allows for solving any predictive problem
  – Using non-parametric solution, at scale

• Unstructured data is fully utilized
  – In-built “feature extractors” for image, text, geo

• Predictive models are built using ALL the available data
  – Overcomes data sparsity challenge using non-parametric methods

• Provenance of predictions explains answer
  – System is not a black-box