On the Practice of Predictive Modeling with Big Data: The Extra Steps that Make the Difference

Nachum Shacham
PayPal

XLDB
5/19/2015
Data Science From 10K ft

- ML Algorithm Design Optimization
- Training set
- Test set(s)
- Model Fitting
- Validate, calculate KPIs
- Scores
- Performance Charts
Predictive Modeling Meets Big Data: Bridging the Gap

Data Science with Big Data: Quantitative change → Process adjustment
Data Exploration: Knowing What’s in the Data and Fitting the Pieces Together

Understand Data Contents & meaning

Distributions and outliers

Null values: codes, quantities

Language

Join keys

Units

Big data visualization

Correlations
Data Munging: Reshaping the Data → Decisions, Decisions

- Dummy variables
- Sub groups
- Categorical variable

- Missing values
- Unify null symbols
- Impute / discard

- Money, time
- Unify values

- Join keys do not match
- Ignore rows

- Irrelevant/redundant values
- Transform
- Nullify
- Discard

- Transform
- Remove outliers

- Skewed distributions
- Standardize

- Reformat keys

© 2015 PayPal Inc. All rights reserved.
Feature Engineering

*Transform raw data to better represent the problem to the model*

**Manual Feature Manipulation**
- Dummy variables
- Text integration
- Sub/super sampling (unbalanced sets)

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Blue</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Green</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Feature learning algorithms**

- Dummy variables
- Text integration
- Sub/super sampling (unbalanced sets)

Integrated in model fitting

Better features → Simpler models, better results
Model Tuning

- Learning Rate
- Interpretability vs Accuracy
- Parameter Grid
- Ensemble: degree, type
Model Evaluation

• Who should judge the value of the results?
  • Evaluation criteria
• What are the cost/value of FP, TP, FN, TP?
  • FP: PR cost
  • FN: Financial loss
  • TP: Intended gain
• Continuous Model evaluation in production
Big Data Machine Learning in A Cloud

- Rich library of ML algorithms
- Auto detect feature format
- Interoperate with ordinary R: split work for complete munging
- Runs through parameter grid
Summary: The “Extra” Steps Make the Difference

Details: the Critical Factor

Platform \(\rightarrow\) Performance

Decisions: understand the impact

Prediction

Delayed
Incomplete
Irrelevant
THANK YOU