• CAPITAL Services Overview
• 20 Dos and Don’ts of Data Science
  – Development
  – Evaluation
  – Implementation/Presentation

- Two Locations:
  - Brookings, SD
  - Sioux Falls, SD

- Manages over $450 MM credit card receivables
- Servicing over 775,000 accounts

- Portfolio Management
- Financial Modeling
- Managed Services
- Compliance
PARD aims to maximize value by moving decision processes along the analytics curve.

Each application progression depends on the incremental expected value gained measured in dollars or customer experience for every hour invested.
Development Tips
Explanation:
• Depending on data size, using the entire data set for data cleaning, report development or model development can slow analysis.

Recommendation:
• Sample data to a reasonable size to develop code/models
• Use the full data once basic design is tested
2 – Incorrectly Assuming Linearity

Prevention Strategy:
- Exploratory Data Analysis
- Transform Variables
  - Weight of evidence binning
- Machine Learning Algorithms
3 - Ignoring Anomalies/Outliers

Profit by Credit Score

Profit
$0 $50 $100 $150 $200

Distribution
0% 5% 10% 15% 20%

Credit Score
600 625 650 675 700 725 750 775 800 825 850

Distribution
Average Profit
3 - Ignoring Anomalies/Outliers

Profit by Credit Score (Binned)

Credit Score

Distribution

Average Profit

Profit

$0 $50 $100 $150 $200

600 625 650 675 700 725 750 775 800+

Distribution

Average Profit
Prevention Strategy:
• Question/investigate significant trends that are outside of expectations
## Predicting Football Score

<table>
<thead>
<tr>
<th>Game</th>
<th>Home Team Score</th>
<th>Previous Wins</th>
<th>Previous Points</th>
<th>Touchdowns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>0</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>2</td>
<td>28</td>
<td>2</td>
</tr>
</tbody>
</table>

## Predicting Customer Defaults

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Loan Default</th>
<th>Credit Score</th>
<th>Previous Defaults</th>
<th>Total Defaults</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>750</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>700</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>675</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>650</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

**Prevention Strategy:**
- Question uncommon good predictor variables
- Review if the information can be systematically used in the future
Description:
- Previous strategies can skew future expectations

Prevention Strategy:
- Reject Inference
- Collect random test data
Example:

- Netflix conducted a predictive modeling contest to better recommend new movies with a $1 Million prize.
- They never used the winning algorithm because of engineering efforts to implement the complicated winner.

Prevention Strategy:

- Be disciplined asking if the complexity is worth the benefits.
- Consider interpretability - Regulators and executives prefer to understand inputs impact on predictions.
Risk Target - KS Chart

Cumulative % Target

Cumulative % Population

- Response Model 1
- Response Model 2
7 – Consider Interactions

Response Model 1 | Response Model 2 | % Difference
--- | --- | ---
Response Rate | 3% | 5% | 67%
Default Rate | 4% | 7% | 75%
Average Profit | $500 | $400 | -20%

Recommendation:
• Assess incremental impact on entire strategy/system
• Multiple Goal Programming
Evaluation Tips
8 - Understand Distribution

Recommendations:
- Always ask what is the distribution of the variable being measured
- Use more summary statistics
  - Variance/Standard Deviation
  - Confidence Intervals
- Show Distribution Chart and Binned Data
8 - Model Comparison

<table>
<thead>
<tr>
<th></th>
<th>KS</th>
<th>AUROC</th>
<th>GINI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.27</td>
<td>0.67</td>
<td>0.34</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.22</td>
<td>0.64</td>
<td>0.28</td>
</tr>
</tbody>
</table>

KS Chart

ROC Chart
9 - Correlation → Causation

**Does employment give people more time to go on Facebook?**

- Facebook Stock
- Employment Rate

96.2% Correlation

---

**Is this tree influencing Detroit's population?**

- Detroit's Population

---

**Does ice cream prevent the flu?**

- Total Flu Patients
- Ice Cream Production

---

**Prevention Strategy:**

- Create randomly controlled tests
- Control for nuisance variables
What caused payment rate to increase?
- New Training?
- Macroeconomic changes?
- New collections strategy?

Need to test to measure actual effectiveness
## Customer Classification

<table>
<thead>
<tr>
<th>Collections Call</th>
<th>Payment</th>
<th>No Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Payment</td>
<td>Do Not Disturb (Sleeping Dog)</td>
<td>Lost Cause (Neutral)</td>
</tr>
<tr>
<td>Payment</td>
<td>Sure Thing (Neutral)</td>
<td>Persuadable</td>
</tr>
</tbody>
</table>
Uplift = Difference = Target Rate_{no dial} – Target Rate_{dial}

Difference in Payment Rate

- Sleeping Dogs: \( \text{Payment Rate}_{no \, dial} > \text{Payment Rate}_{dial} \)
- Lost Causes/Sure Things: \( \text{Payment Rate}_{no \, dial} \approx \text{Payment Rate}_{dial} \)
- Persuadables: \( \text{Payment Rate}_{no \, dial} < \text{Payment Rate}_{dial} \)
Description:
• Testing multiple hypotheses can lead to false positives

Prevention Strategy:
• Bonferroni Correction (simple but weak power)
• Cross validation and ongoing/iterative testing

<table>
<thead>
<tr>
<th>Marketing Test</th>
<th>Control</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Response Rate</td>
<td>2.0%</td>
<td>2.1%</td>
<td>2.3%</td>
<td>2.4%</td>
<td>3.2%</td>
</tr>
<tr>
<td>P-Value</td>
<td>-</td>
<td>0.436</td>
<td>0.323</td>
<td>0.184</td>
<td>0.046</td>
</tr>
</tbody>
</table>
### KS Statistics

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>...</th>
<th>Model 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.74</td>
<td>0.75</td>
<td>0.73</td>
<td>0.72</td>
<td>...</td>
<td>0.65</td>
</tr>
<tr>
<td>Validation</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
<td>0.70</td>
<td>...</td>
<td>0.59</td>
</tr>
</tbody>
</table>

- Model 1 is declared the winner since it has the highest validation KS Statistic.
### KS Statistics

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>...</th>
<th>Model 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>0.74</td>
<td>0.75</td>
<td>0.73</td>
<td>0.72</td>
<td>...</td>
<td>0.65</td>
</tr>
<tr>
<td>Validation</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
<td>0.70</td>
<td>...</td>
<td>0.59</td>
</tr>
<tr>
<td>Test</td>
<td>0.61</td>
<td>0.65</td>
<td>0.71</td>
<td>0.65</td>
<td>...</td>
<td>0.61</td>
</tr>
</tbody>
</table>

- Validation data becomes less generalizable to future application as the number of models tested increases.
- Additional test data sets can help measure uncertainty of roll out.
Default Rate by Change in Credit Score

Change in Credit Score

Default Rate

0% 2% 4% 6% 8% 10% 12%

-20 -15 -10 -5 0 5 10 15 20
Description:
- Hindsight bias can lead to overconfidence and errors in decision making.

Prevention Strategy:
- Think about how alternative hypotheses can be correct
- Ongoing testing
Description:
• Too short of performance horizons can lead to incorrect conclusions

Prevention Strategy:
• Create reliable models to forecast long term impacts
• Continue to monitor tests if there is a possibility of performance shifts.
Description:
• Validation is usually complete after a model is implemented
• Ongoing validation is critical for models that have a high impact on business
• Ongoing validation measures changes from
  – Macroeconomic environment
  – Competition
  – Updated input variables
  – Policy changes
Implementations & Presentation Tips
• Automation Pros:
  – Long term time savings
  – Error mitigation

• Automation Cons:
  – Short term costs including opportunity costs
  – Future changes take more effort
Description:
• Over precision can lead to over confidence of predictions.

Prevention Strategy:
• Communicate confidence using intervals
• List assumptions of predictions
• Simulate “what if” scenarios to show variability
Description:
• Overvaluing past investments even if future is gloomy.

Prevention Strategy:
• Look at future expectations independent of past investments
• Know when to fold’em
Description:
- Confirmation bias leads to higher probabilities of conclusions inline with past beliefs.

Prevention Strategy:
- Set decision criteria before tests
- Have independent analyst validate
- Stay objective
Description:
• Standard prediction intervals include statistical uncertainty.
• Considering structural uncertainty widens intervals closer to real world outcomes.

Prevention Strategy:
• Test how models perform in simulated rollout scenarios
• Consider changes over time for model predictions
• Consider omitted variable bias
Recommendation:

• Have a continuous learning mindset in your career
  – Go to conferences
  – Explore new data (Kaggle/CrowdANALYTIX)
  – Research with interests you

• Have a iterative mindset during projects
  – New question arise during analysis and answers to these questions often have bigger impacts than the original question
  – Don’t be satisfied with the first conclusion
  – Important problems rarely have straight forward linear solutions
Description:
- A data scientist could build the best model for the perfect application. If stakeholders don’t understand the proposal and don’t implement the value = zero.
- Success only happens when work is implemented

Recommendation:
- Invest in building trust and an analytics culture
- Involve stakeholders early and often
- Use the power of stories
Recap

• Stay objective
  – Ongoing Testing
  – Independent reviews
  – Question assumptions

• Don’t lose sight of end goal
  – Balance complexity with benefits
  – Consider tradeoffs between metrics
  – Define appropriate long term goals