Summary of AI in Marketing

Michael Thomas

AI in Marketing - Semester II, 2022

Big Picture

- Innovations in prediction technology allows for new opportunities to monetize predictions.
- In marketing applications, this primarily consists of customizing marketing to each individual.
  - Each individual belongs to their own segment.
  - Algorithms can automate delivery of marketing content suited to individual needs.

Business Implications

- Shortage of people knowledgeable of machine learning presents opportunities.
- Data allow for better predictions. Big companies make big plays to acquire more data.
- Winner take all competitive environment.
  - It's hard to compete with a a business with more data.
  - Their predictions, offers, customization will be better.

The Value of Predictions

Predictions can make more money by improving efficiencies

- Who will respond to an ad?
- Who will churn?
- Who will donate the most?
- Who will buy the most?
- Who will respond to a sales call?

Tends to work best when you are predicting low-probability events.

Binary Outcomes

We can capture the economics of these outcomes with a **cost-benefit matrix**

	Predicted 0	Predicted 1
Actual 0	$v(o_1)$	$v(o_3)$
Actual 1	$v(o_2)$	$v(o_4)$

- Usually $v(o_1)$ and $v(o_2)$ are zero.
- $v(o_3)$ is the cost of your marketing instrument when the customer does not respond
- $v(o_4)$ is the net value of correctly predicting which customer will respond.

Binary Outcomes

We can capture the probabilities of correct and incorrect predictions with the **confusion matrix**

	Predicted 0	Predicted 1
Actual 0	$p(o_1)$	$p(o_3)$
Actual 1	$p(o_2)$	$p(o_4)$

- Can be estimated from a holdout sample.
- Compare your predictions to the actual outcomes.
- Requires a threshold for predicting someone will respond.
- The threshold can be adjusted to optimize profits.

Calculate Expected Value for Binary Outcomes

Confusion Matrix Probabilities

	Predicted 0	Predicted 1
Actual 0	$p(o_1) = 0.57$	$p(o_3) = 0.14$
Actual 1	$p(o_2) = 0.21$	$p(o_4) = 0.13$

Cost-Benefit Matrix (using Problem Settings 3 here)

Cost-Benefit matrix relative to doing nothing to prevent churn

	Predicted 0	Predicted 1
Actual 0	$v(o_1) = \$0$	$v(o_3) = -\$1$
Actual 1	$v(o_2) = \$0$	$v(o_4) = \$10 - \$1 = \$9$

Expected Value (EV) of the Model's Predictions

$EV = p(o_1) \times v(o_1) + p(o_2) \times v(o_2) + p(o_3) \times v(o_3) + p(o_4) \times v(o_4)$

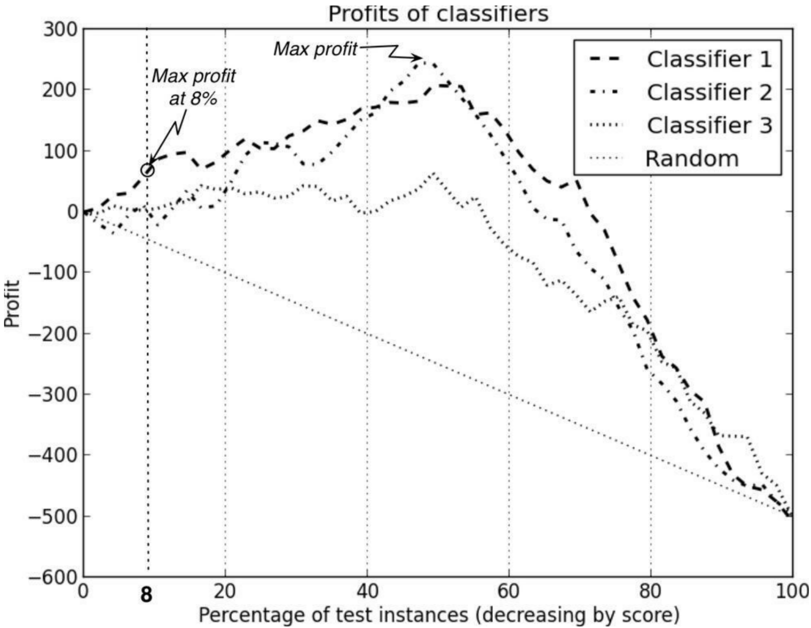
$EV = 0.57 \times \$0 + 0.21 \times \$0 + 0.14 \times (-\$1) + 0.13 \times \$9$

$EV = \$1.03 =$ Average profit per customer from churn targeting.

Scoring

- Predictive models allow us to **score** our customers
  - Rank them from most attractive to least attractive.
- The profit model allows us to decide how many customers to target
  - The most attractive customers may be profitable.
  - Less attractive customers will cost more than they are worth.

Profit curve



Overfit

- Using a model that is too flexible for the data
- Picks up noisy features of the data that do not generalize.
- Core concept and problem in machine learning.
  - Algorithms attempt to balance overfit and underfit for optimal predictions
- Different methods exist to manage overfit.

Cross Validation

- Use the same data set to train and validate predictions repeatedly.
- Divide the data into random folds.
- Each fold gets a chance to be the "test set."
- Gives multiple estimates of model prediction. Can average across these.

LASSO

- Take OLS and add a penalty for having coefficients different from zero
- Selects which variables to include in the model.
- Shrinks variables included in the model toward zero.

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k \right)^2 + \lambda \sum_{k=1}^{p} | \beta_k |$$

- Key tuning parameter: λ.
  - $\lambda = 0$ is the same as OLS
  - $\lambda \to \infty$ only the intercept is included in the model
  - Use cross validation to discover the optimal λ.

Trees

- Trees can describe decision making processes.
- Also, **regression trees** allow multiple variable to work together to predict an outcome.
  - At each node, the algorithm looks for cut points to divide data.
  - Whichever variable has the "best" cut point divides the data at that node.
  - Repeat.
- Essentially, treats different subsets of X as having common outcomes.
- Pro: Allow for lots of flexibility and interactions among the X variables.
- Con: Tend to overfit, especially if allowed to grow too deep.

Random Forests

- Take the best features of trees and improve on them.
- **Ensemble** of trees – many trees estimated together.
  - Each tree is estimated on a different **bootstrapped** data set.
  - Each node can be split using a random subset of features
    - "feature bagging"
  - Injecting randomization into the algorithm (through bootstrapping and feature bagging) helps improve predictions.
  - Each tree makes a prediction.
  - Average across all those predictions for the Random Forest prediction.

Random Forests

- Pros:
  - Flexible, non-parametric estimates.
  - Can approximate continuous functions.
  - Little tuning required, typically.
- Cons:
  - Takes a long time to train.
  - Requires more data than linear models.
  - Including non-predictive features hurts its performance.
  - Difficult to interpret what drove the predictions.
  - Gradient boosting methods now often do better, if tuned.
    - E.g., XGBoost

Experiments

- Experiments generate unbiased, causal estimates.
- Experiments work through randomization to create two samples that are
  - Identical in expectation
  - Differ only by treatment assignment.
- See limited use in business because:
  - Can be expensive to run.
  - Lack of infrastructure and understanding of their value.
- Causal estimates are usually what marketers want:
  - What happens if I change marketing instrument, X?

Heterogeneous Treatment Effects

- Heterogeneous treatment effects refer to how different groups respond differently to treatment.
- Also a core question for marketing:
  - **Who** will respond best to marketing instrument, X?
- Regressions on different populations in the data provide estimates of heterogeneous treatment effects.
- Machine Learning can be combined with experimental data to find heterogeneous treatment effects.
  - E.g., Causal Forest.

Multi-Armed Bandits

- Method to optimize use of among multiple versions of an ad.
- Starts by **experimenting** to discover which version of the ad might work best (e.g., highest click-through rate.)
- Proceeds to **exploiting** the results from these experiments by relying on the best performer.
- **Thompson Sampling** provides a simple heuristic for a smooth transition from explore to exploit.
  - Relies on Bayesian statistics to characterize beliefs on the CTR for each ad.