An Overview of Machine Learning

Presented by Hunter Boles

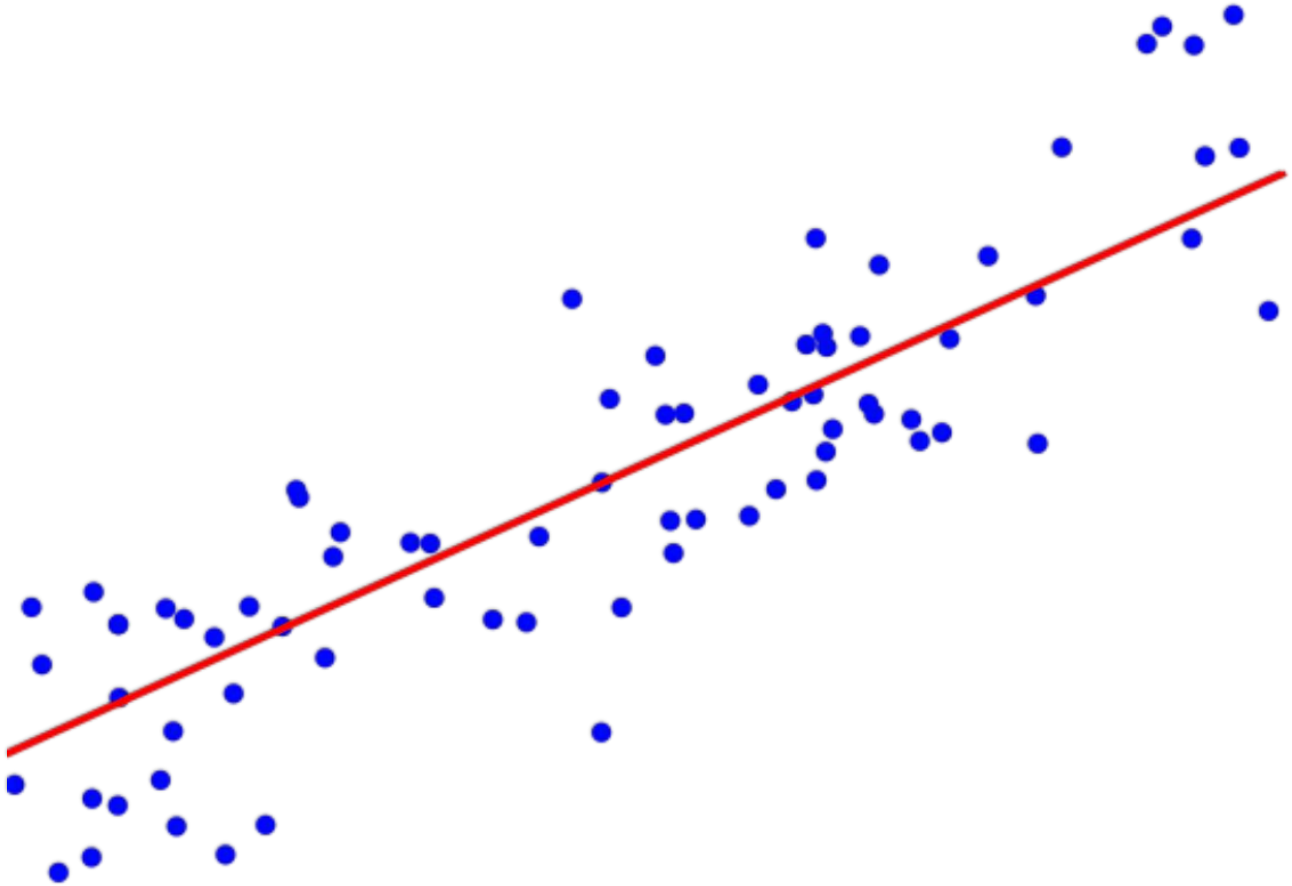
Overview

- Job Opportunities
- What is ML?
- Valuable Lessons for ML
- Case Study: Probability Calibration

Job Opportunities

- Data Analyst (CS/Statistics)
  - Analyze what has happened in the past
- Data Engineer (CS)
  - Prepare the data for ML models
- Data Scientist (CS/Statistics)
  - Build an ML model
- Machine Learning Engineer (CS)
  - Optimize the ML model for a production setting.

What is Machine Learning?



Machine Learning

- “Computer algorithms that can improve automatically through experience and by the use of data” – Wikipedia
- Considered a subset of Artificial Intelligence.
  - The AI that is not considered ML is normally rewards-driven and has a clear agent and transition states.
  - ML is still optimizing to a given metric but does not require an agent to do so.
    - I.E. Linear Regression

What can Machine Learning Solve?

- Regression problems
  - Given data, predict a value.
  - How much will my house be worth in 6 months?
- Classification problems
  - Predicts a class (or probability).
  - Can we predict who will fall asleep during this presentation?
- Clustering problems
  - Predicts class membership when we don’t know the classes during training
  - Which of these doesn’t belong?

Typical ML Model Development

- Determine problem statement (and determine if it is classification/regression/clustering)
- Clean data (This is 80% of the work)
- Train model
- Review (This is the other 15%)

What should data look like for ML models?

	A	B	C
1	**Month**	**Rainfall (mm)**	**Umbrellas sold**
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37

Lessons Learned

Definitions

- Definitions are single-handedly the most important part of data understanding.
  - Where is the data from?
  - What was done to the data?
  - What are the assumptions?
- Example: Predicting when a loan goes derogatory.
  - What is a loan (and how is it represented)?
  - What does it mean to be derogatory?

Definitions (Cont'd)

- What happens when there are not clear definitions?
  - Miscommunication
  - Subtleties in the problem are missed
  - Unable to predict what the business is truly interested in
- Metrics

Business Value

- How models improve the business is another key aspect for ML (And often forgotten about)
- Applying a kitchen-sink approach without understanding the variables is not smart
  - I.E. Loan servicer data commonly contains zip code information. When considering underwriting, zip code can be a proxy for race (and using race in underwriting is illegal!). Keep the business use-case in mind when choosing variables

Communication

- Technical communication is, unfortunately, important.
  - Ties together the last 2 points as well.
- Communicate models, assumptions to stakeholders in non-technical way.

Aside: Why Data Scientists are still important

- Many fear that AI will overtake a Data Scientist's position in an organization.
  - Many AutoML libraries already exist.
- Data scientists are needed to determine features of interest, business value, and communicate with shareholders.

Probability Calibration

(For classification models)

The Problem

- The model predicts a certain probability when it truly occurs at another.
- Example: Weather Forecasting
  - Weatherman says there is a 95% chance of it raining every day.
  - … And it never seems to rain. (Maybe once every 3 weeks)

Implications

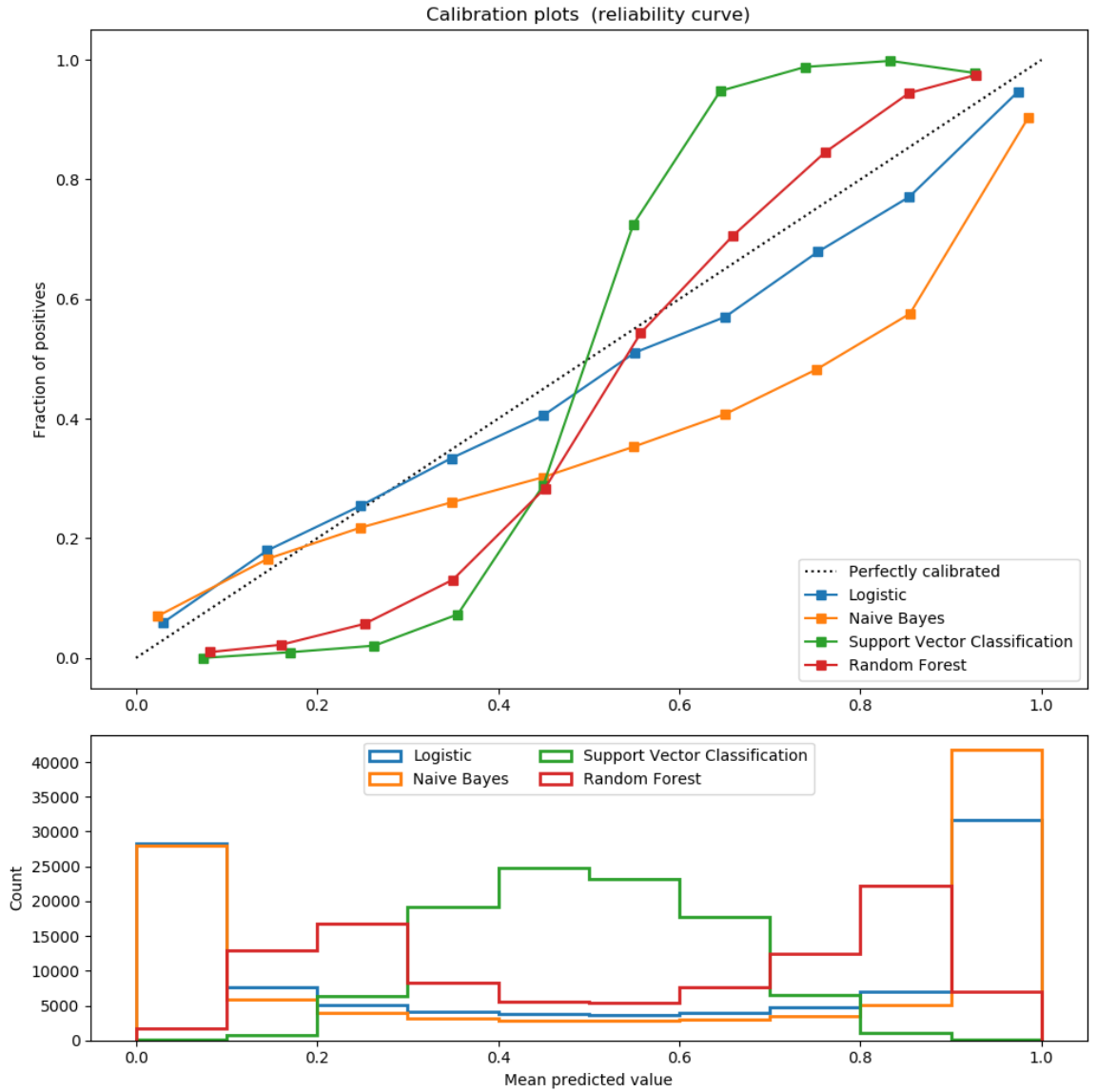
- An event seems more/less rare than it truly is.
- Becomes an issue when decisions are made around the probability of an event happening.

What causes it?

- Over/underweighting a certain class.
  - This is needed for models that have little to no data for a certain class, like derogatory loans.
- Using a function other than log loss to optimize probabilities.
  - Logistic regression uses this
  - Decision trees don’t use log loss.

What Uncalibrated Probabilities look like

- The training population is binned into *n* buckets.
- The fraction of true positives is given on the y axis per bucket
- The x-axis is the mean predicted probability of each bucket.



Solutions

- Determine if class weighting/model type is needed for good results.
- Use a calibrator

Calibrators

- Finds a mapping of uncalibrated probabilities to calibrated probabilities
- Main solutions:
  - Isotonic Regression
  - Platt Scaling

Platt Scaling

- Fit a univariate logistic regression model between predicted and true probabilities.

Isotonic Regression

- Fit a non-linear, nondecreasing piecewise function to the predicted probabilities versus actual
- Works better than Platt scaling when large datasets are available