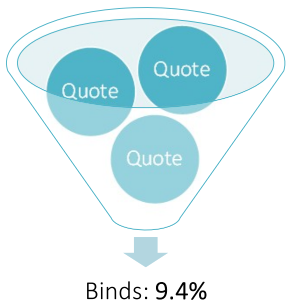
Proposal: Close Rate Prediction Model

Joe Warth

Business Problem & Use Cases

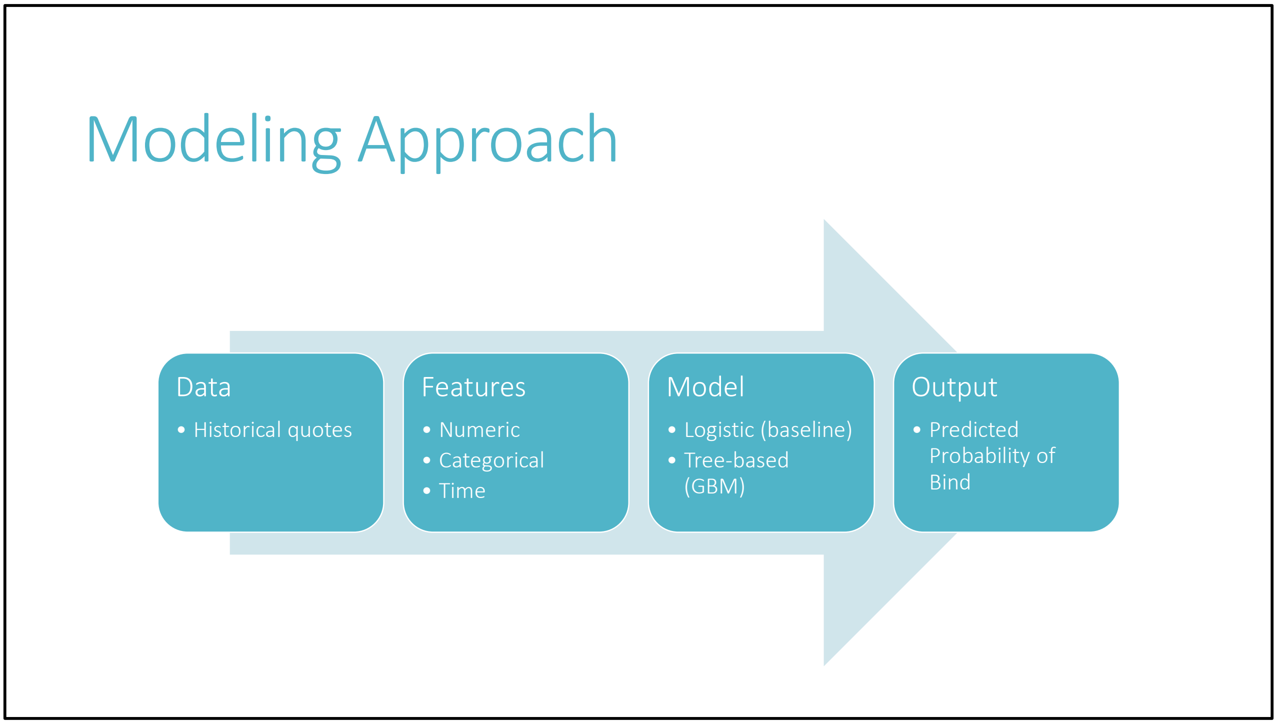


Prioritize high-likelihood quotes

Find low-hit rate segments

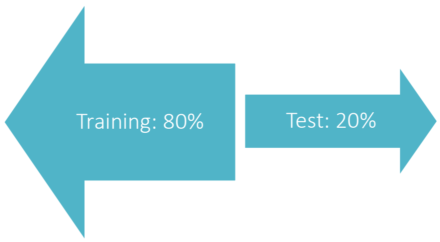
What-if pricing & strategy

- We get many quotes
- Only about 9.4% of those quotes ultimately bind into a policy
- In that process, **many** quotes are manually touched by underwriters, even if they have very little chance of becoming a policy
- Imagine if we could prioritize high-likelihood quotes in the underwriting process while deprioritizing low-likelihood quotes
- For our product and pricing colleagues, it would be nice to highlight segments with ultra-high or ultra-low likelihood of binding to help them fine tune their pricing to fit the market better than our past loss cost experience alone can



- Our proposal is to use our historical quotes and develop numeric, categorical and time-based features to feed into a model
- The model would be a binary classification model
  - There are many approaches to use for this task, but we recommend starting with a simple logistic regression model as a baseline
  - This baseline model will be highly interpretable and easy to understand,
  - But we may get a better fit out of a more sophisticated tree-based model like a GBM because of its ability to handle nonlinearity and interaction terms more robustly
- Our output from these models would be a predicted probability of bind, which we can then use to solve the problems we discussed on the prior slide

Train/Test Strategy



- Test model on most recent ~20% of quotes as a “future-like” period
- Optimize model for future prediction rather than past description

Evaluation Metrics		
ROC AUC	Brier Score	Calibration curves

Time-based split, not random:

  Train: older months

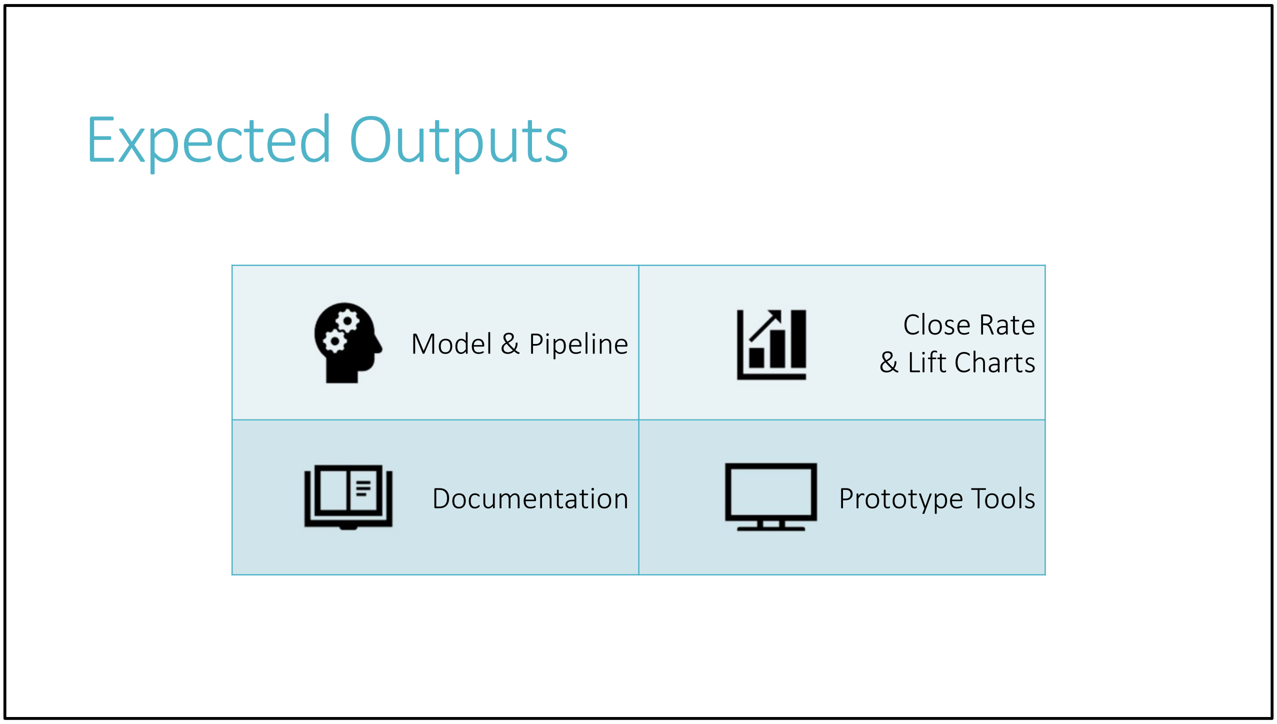
  Test: most recent 20% of rows

Ensures we evaluate, “Can this model predict future quotes?”

Key metrics:

  Discrimination: ROC AUC

  Probability quality: Brier score, calibration curves



Model artifacts:

- Trained model + feature pipeline (code + saved objects)

Summary tables and plots:

- Overall and segment-level hit rate vs predicted
- Calibration and lift charts

Documentation:

- Data prep & feature definitions
- Model limitations and appropriate use

Prototype Tools:

- Model-based pricing recommendations
- Score a batch of quotes & visualize predicted probabilities of bind

Assumptions, Risks & Limitations

Assumptions	Risks & Limitations
★ Data is representative ★ Key drivers are captured ★ Stable quoting process	★ Small dataset ★ Market/pricing drift over time ★ Probabilities are estimates with error, not guarantees

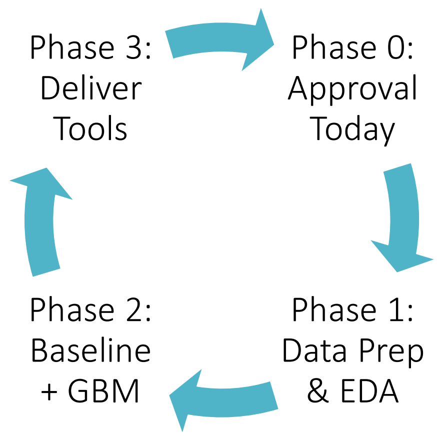
Assumptions

- Historical data is reasonably representative of current quoting environment
- Key drivers of hit rate are captured in the available fields

Risks/limitations

- Data size: current sample ~4.3K quotes, only ~9% are binds
- Drift: model performance may degrade if pricing/market changes substantially
- Interpretation: probabilities are estimates, not guarantees; should inform, not dictate, decisions

Timeline & Next Steps



Phase 1: Data prep & EDA
- Cleaning, target definition, stakeholder signoff

Phase 2: Baseline modeling
- Logistic regression
- Tree-based model

Phase 3: Decisions & Delivery
- Select final model
- Deliver tools for utilizing the model

In order to detect and handle model drift and always-changing markets, this process would need to be iterative with a regular update cadence

Decision needed now: Approval to proceed with Phase 1 and access any additional fields/data sources that could materially improve the model (e.g., competitive position, agent attributes, prior relationship indicators, etc.)