INSURANCE, MY DEAR DATA
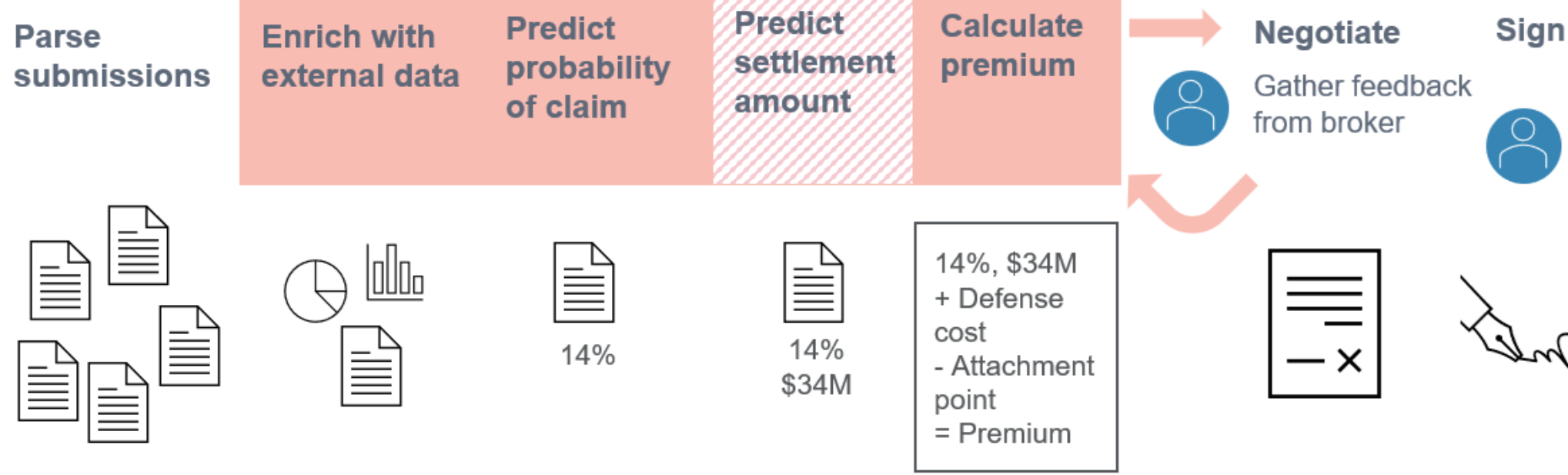## PREDICTING WHO WILL GET SUED

Moritz Bartusch & Megi Jaupi
Faculty Advisor: Rahul Mazumder   **Banyan Advisor:** Michael Densham

---

BANYAN'S VISION

**Banyan strives to change a century old industry by building a data-driven underwriting process**

Capstone project

Parse submissions — Enrich with external data — Predict probability of claim — Predict settlement amount — Calculate premium

14% → 14% $34M → 14%, $34M + Defense cost - Attachment point = Premium

Negotiate — Gather feedback from broker — Sign

PROBLEM STATEMENT

Develop a model to predict claims and tools to help the vision

(?) Is the industry looking at the right elements for prediction?

(?) Can we predict claims with a small number of factors?

(?) What are the factors that matter?
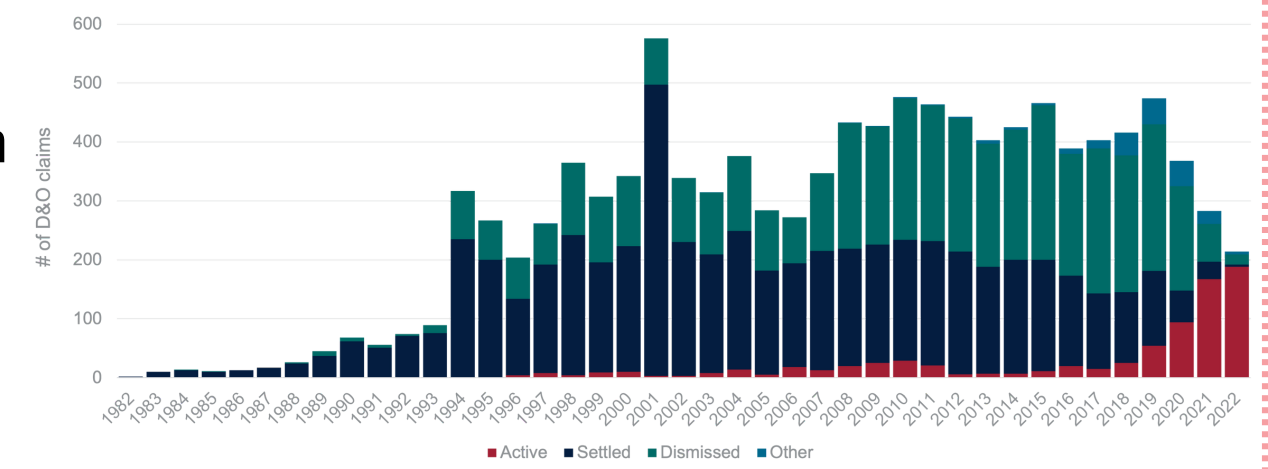
FACTOR CONSTRUCTION

Using academic research and Banyan's experience we formulated predictive hypothesis for claims

Size comes with responsibilities
Once a liar, always a liar
Some owners have more litigations

DATA EXTRACTION PIPELINE

Developed automated data extraction pipelines from several sources: CapIQ, ISS. Created a methodology to merge companies with securities, clean, enrich the data.

What is the market capitalization?
Was the company sued before?
Who is the owner?

FEATURE ENGINEERING

Engineered transformations and variations of factor measures to construct predictive features.

Calculate the **growth** of market capitalization.
Extract the **result** of the litigation
Capture **changes** in owners

MODELLING

HIGH DIMENSIONALITY
+8000 US companies
+100 Features
10 Years of data per company per feature

MODELS
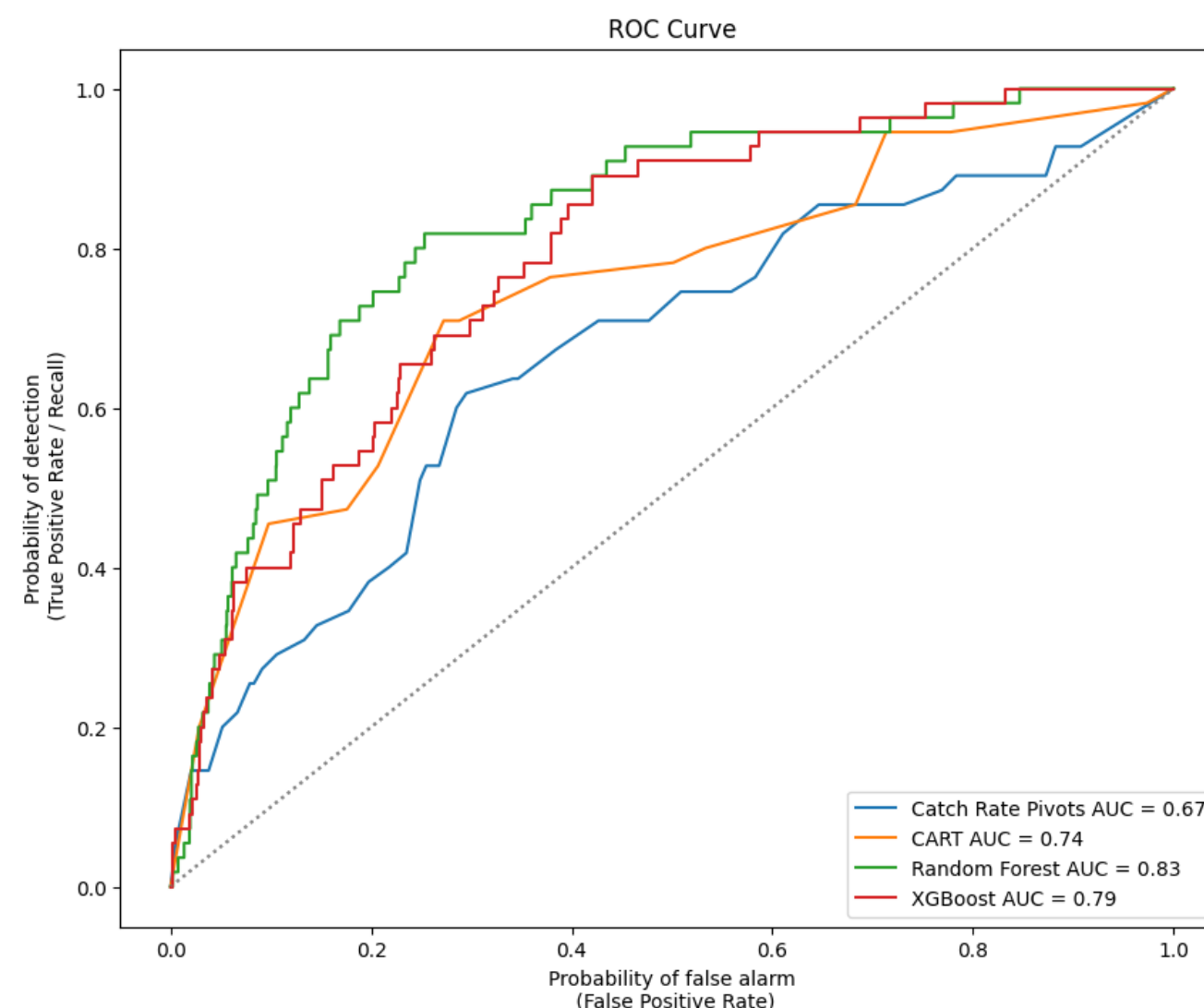CART
XGBOOST
RANDOM FORESTS

CLASS IMBALANCE
SMOTE

GLOBALISATION
3000 more companies for Canada

RESULTS

+16%

Increase in AUC our best model versus the baseline

ROC Curve — Catch Rate Pivots AUC = 0.67, CART AUC = 0.74, Random Forest AUC = 0.83, XGBoost AUC = 0.79

IMPACT

+3h saved per submission

+$9,000,000 in premium per person

(target) The insurance industry is not looking at all the relevant factors - rigid view on market cap and sector

(target) For large companies: size, short interest, volatility, employee growth, institutional ownership matter

(target) For small companies: size, healthcare sector, financial sector, quality of earnings matter

---

WEB APP

Search any public company in the US by its ticker

Average 3Y Market Cap <= 71.479.426
Claim ratio: 411 / 16918
Probability of Claim: 2.43%

Mrket Cap Relative Volatility <= 0.585
Claim ratio: 364 / 16418
Probability of Claim: 2.22%

Claim ratio: 47 / 500
Probability of Claim: 9.40%

Hist Volatility of Price <= 44.677
Claim ratio: 263 / 14107
Probability of Claim: 1.86%

Short Interest Percent <= 1.118
Claim ratio: 101 / 2311
Probability of Claim: 4.37%

Claim ratio: 13 / 810
Probability of Claim: 1.60%

Institutional Ownership Percent
Claim ratio: 88 / 1501
Probability of Claim: 5.86%

Settlement Rate	30.51%
Number of claim company-years	92
Active	33
Settled	18
Severity	$9.25 Mio.
Median Settlement	$14.25 Mio.
Defense Cost (Mio.)	10
Attachment Point	$15 Mio.

Number of Cases by Settlement Amount

Implied Loss — $0.17 Mio.

COMPANY   PEER OUTCOMES   LIST OF CLAIMS