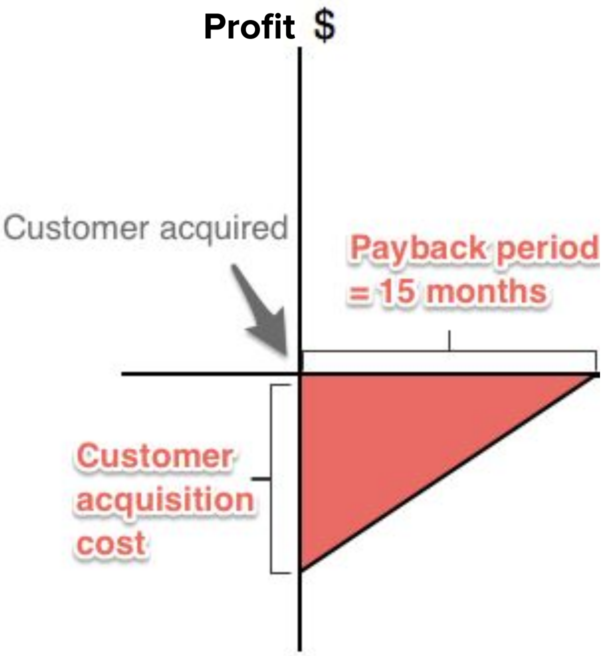
Predicting Cross-Sold Customers

Andrew Smith

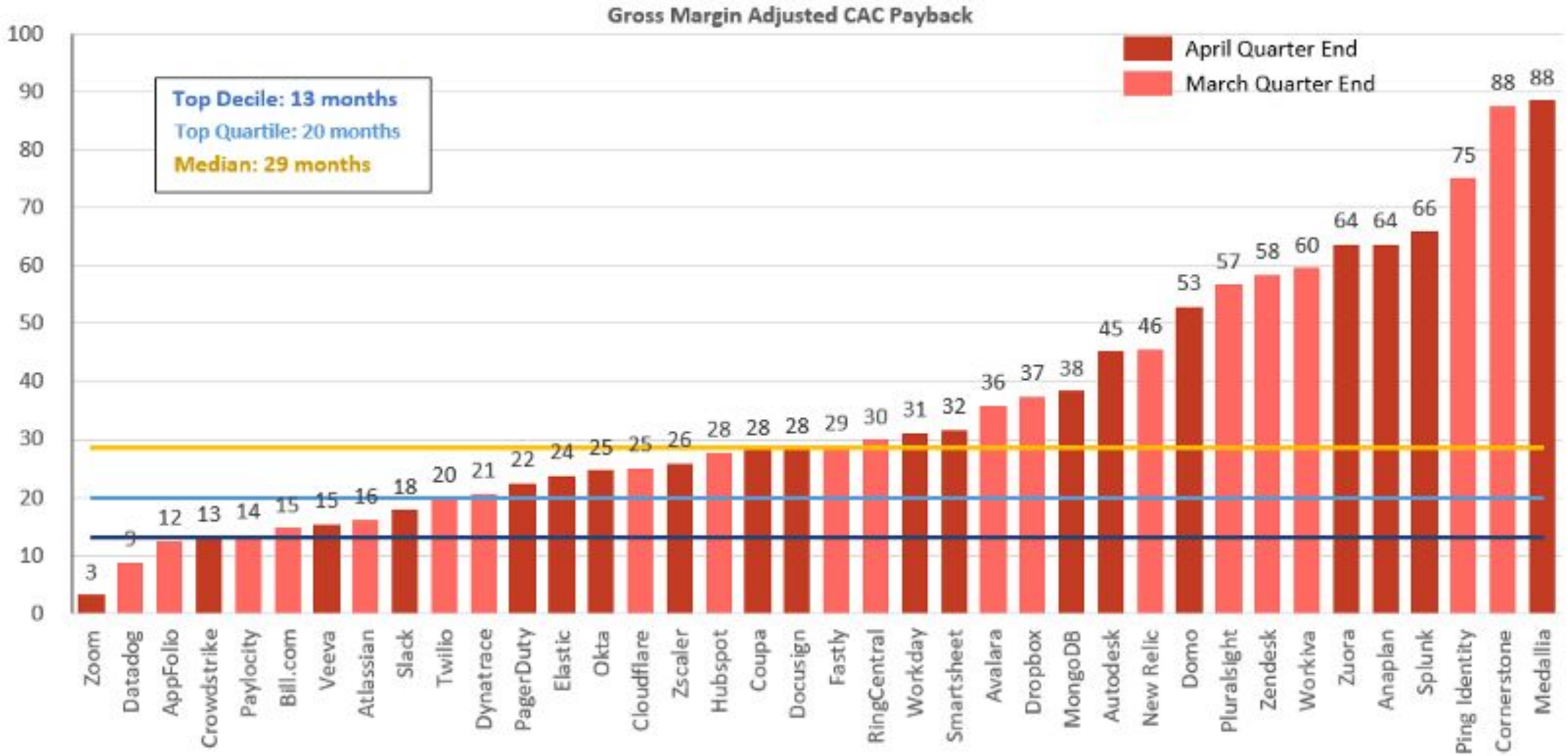
Goals

- Predict whether an existing health insurance customer will be cross-sold car insurance

Cross-Selling = Profitable Growth



Payback Distribution



Use Case

- Give special attention to these customers to try to ensure they are cross-sold and ignore customers with a low probability of conversion
    - Dynamic Pricing (discounts to select customers)
    - Pushed Marketing Campaigns
    - More attention by the sales team

- Key Metric of Interest:
    - F2 Score - places less weight on precision with more weight on recall
    - Having a wider funnel of cross-sell customers is best

Tools

Modelling / Cleaning / Viz

pandas

scikit learn

Imbalanced Learn

seaborn

Data Sources & Data Description

- Data sources:
    - Kaggle Dataset, 381,109 rows
- Key variables:
    - Labels:
        - 0 (not cross-sold) / 1 (cross-sold)
    - Features *(9)*:
        - Gender
        - Age
        - Drivers License (Y/N)
        - Region Code
        - Previously Insured (Y/N)
        - Vehicle Age
        - Annual Premium
        - Policy Sales Channel
        - Vintage (Days as customer)

Model Workflow

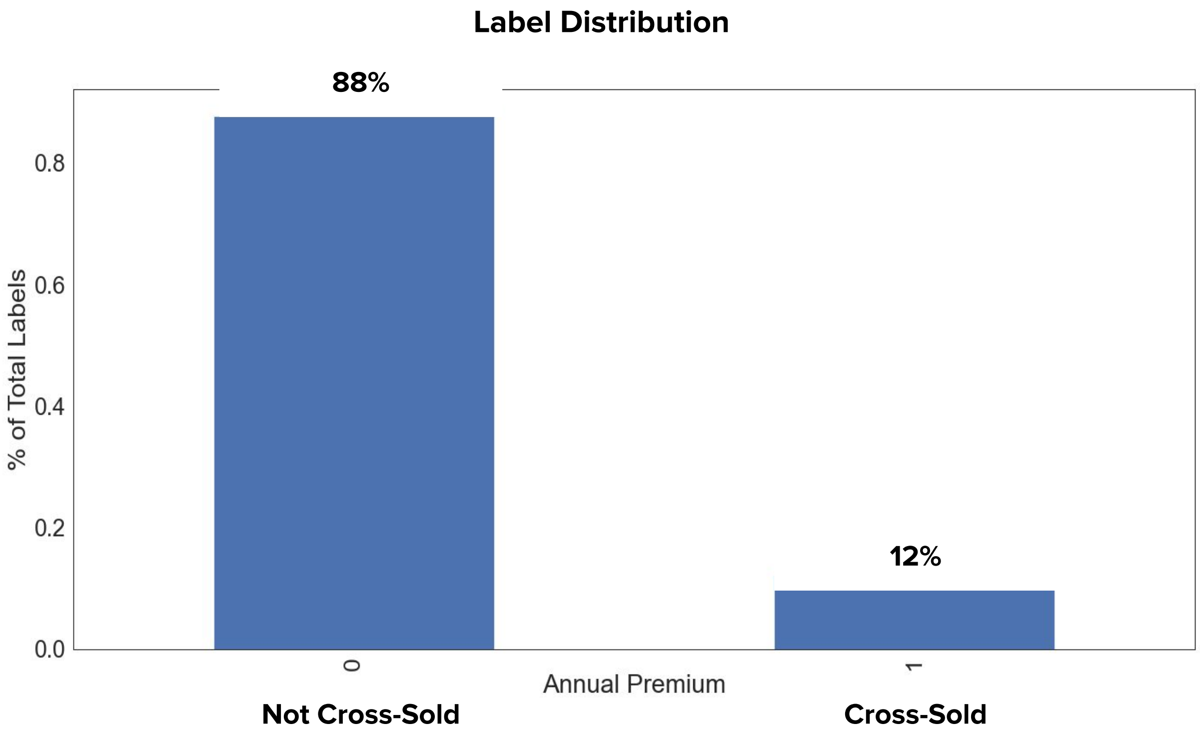
1. Baseline Models
2. Sampling Methods on Baseline Models w/ Parameter tuning
3. Feature Engineering
4. Random Bagging Methods

Baseline Models

Model	F2	Accuracy	Precision	Recall	Notes
Logistic Regression	61.0%	75.4%	30.8%	80.7%	Threshold of 0.2
Categorical Naiive Bayes	38.0%	78.3%	29.4%	53.7%	Categorical Only
KNN (n_neighbors = 5)	16.7%	85.3%	30.2%	15.0%	Various K's tested
Random Forest	8.7%	87.2%	37.8%	7.3%	--
XG Boost (Binary Logistic)	0.7%	87.8%	53.1%	0.6%	Optimizied for error
Gaussian Naiive Bayes	0.4%	87.7%	25.5%	0.3%	Continuous Only

** Represent test Scores after a train, test, val split*

Baseline Models



Model Workflow

1. Baseline Models
2. Sampling Methods on Baseline Models w/ Parameter tuning
3. Feature Engineering
4. Random Bagging Methods

Baseline Models w/ Imbalance Sampling and Hyperparameter Tuning

Model	F2	F1	Accuracy	Precision	Recall	Notes
Logistic Regression	60.5%	44.9%	76.3%	31.4%	78.9%	Oversampled, Threshold of 0.65
KNN (n_neighbors = 29)	59.3%	40.4%	68.8%	26.4%	86.1%	Oversampled, K optimized, feature engineered
Categorical Naiive Bayes	45.7%	29.7%	58.8%	18.8%	71.2%	Oversampled, Categorical Only
XG Boost (Binary Logistic)	45.3%	39.7%	81.4%	33.0%	49.9%	Scale_pos_weight optimized
Gaussian Naiive Bayes	45.7%	37.7%	78.6%	29.2%	53.3%	Oversampled, Continuous Only
Random Forest	24.1%	27.5%	85.6%	36.0%	22.2%	Oversampled

** Represent test Scores after a train, test, val split*

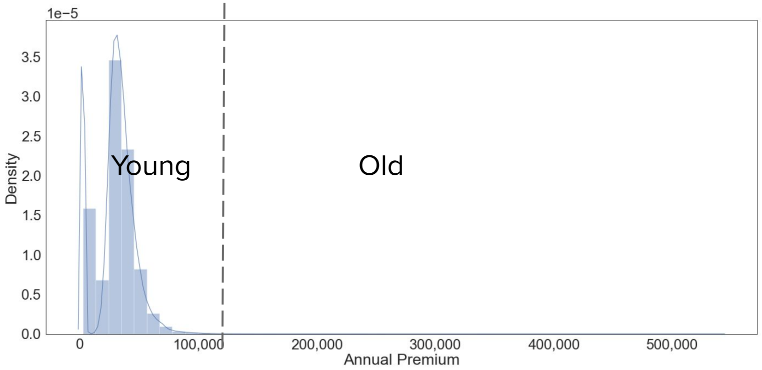
Model Workflow

1. Baseline Models
2. Sampling Methods on Baseline Models w/ Parameter tuning
3. Feature Engineering
4. Random Bagging Methods

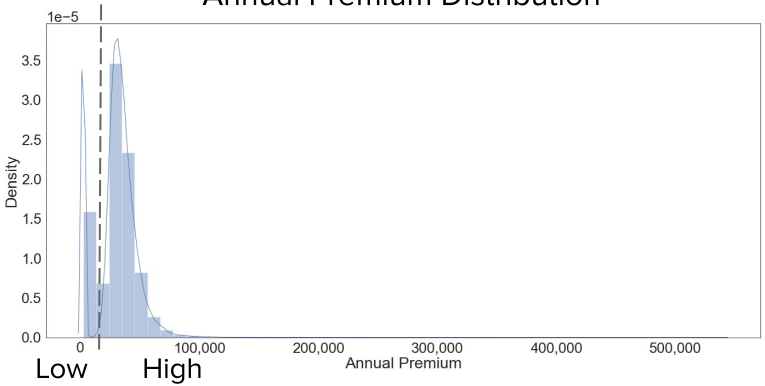
Feature Engineering

Categorical Variables

Age Distribution



Annual Premium Distribution



Interaction Terms

Age **X** Vintage
Young **X** Large Premium
Vehicle Damage **X** Vehicle Age
Gender **X** Age

No impact on top model (logistic regression)

Model Workflow

1. Baseline Models
2. Sampling Methods on Baseline Models w/ Parameter tuning
3. Feature Engineering
4. Random Bagging Methods

Undersampling Bagging Methods

Model	F2	F1	Accuracy	Precision	Recall	Notes
Balanced Random Forest	60.2%	42.8%	73.0%	28.9%	82.5%	Undersampling bagging
Balanced Bagging Classifier	56.0%	42.4%	76.3%	30.2%	71.3%	Undersampling bagging

** Represent test Scores after a train, test, val split*

Top Model: Logistic Regression *(threshold = 0.2)*

61%	**81%**	**75%**
F2 Score	Recall	Accuracy

Confusion Matrix

True Positives (Predicted no cross-sell and actual is no cross-sell) **498,896**	**False Negatives** (Predicted cross-sell and actual is no cross-sell) **16,984**
False Positives (Predicted no cross-sell and actual is cross-sell) **1,701**	**True Negatives** (Predicted cross-sell and actual is cross-sell) **7,641**

Top Features

Vehicle Damage
2.64

Policy Sales Channel 26
1.45

Region Code 28
1.44

Model Weakness / Next Steps

- Low precision
- Request further features
- Speak with domain experts to brainstorm further features
- Try ensembling

Questions
?

Appendix: Feature Pairplot