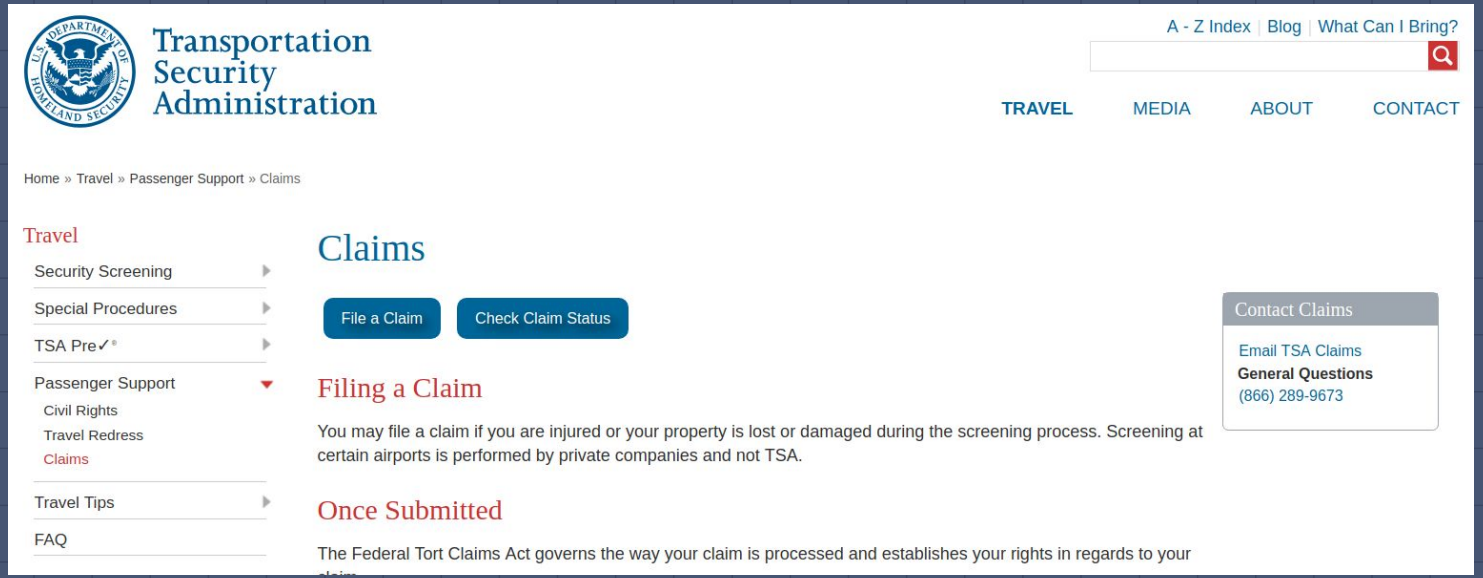
Predicting TSA Claim Results

Perry Chu

What Are TSA Claims?

1) Airport
2) Injury / damage due to security screening
3) https://www.tsa.gov/travel/passenger-support/claims



Why is it useful?

- Passenger (or Travel Insurance)
  - Is it worth the taking time to file?
  - Will TSA reimburse this claim?

- TSA:
  - How should we triage new claims?
  - Which claims should we prevent to reduce cost?

Data Set

- Kaggle Dataset
  - 200k published claims from 2002-2015
  - Source: www.dhs.gov/tsa-claims-data

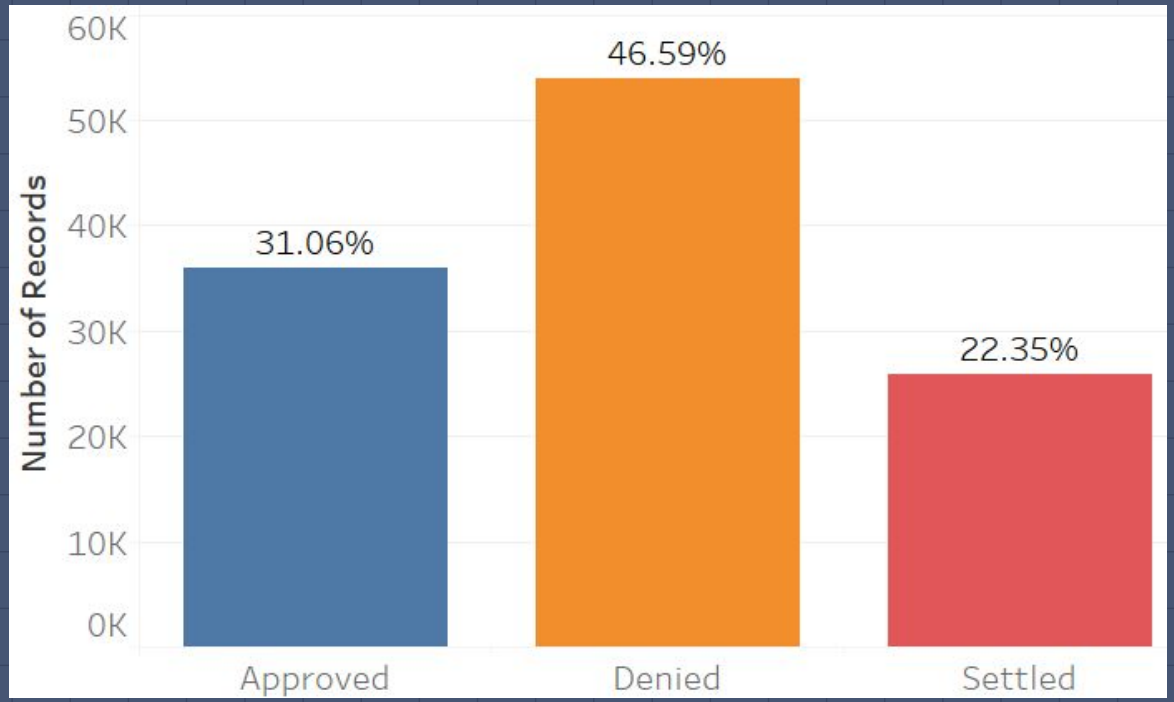
- Columns:

Claim Result	Claim Type	*Item Description*
Claim Amount	Claim Site	*Disposition*
Report Date	Airport Code / Name	*Close Amount*
Incident Date	Airline	*Claim ID*

Target
Numeric
Categorical
Discarded

Target Class Balance

Claim Status



Workflow

1. Data Cleaning
2. Feature Engineering
3. Feature Selection
4. Model Selection
5. Model Tuning

Workflow - Data Processing

Data Cleaning

Removed rows missing > 5 values

Drop nulls for numeric columns

Filled nulls for categorical columns

Fixed:

- Date formats
- Number formats
- Misspelled / inconsistent category values

~140k remaining rows

Feature Engineering

Generalize dates

- Delay (report - incident)
- Day, Month, ~~Year~~

Merge low frequency categories

Convert categorical to numeric

- Dummy vars
- Frequency (in train set)
  - Rank
  - Count

~18 Features

Workflow - Selection / Validation

Feature Selection

[Model: Random Forest]

	Accuracy
Baseline (Claim Amount)	51.5
+Date	52.7
+Category	
▫ Dummy	53.7
▫ Rank	54.2
▫ **Count**	**54.2**
▫ Rank + Count	54.1
+ Date + Count	**56.5**

Model Selection

[Features: Date + Count]

	Accuracy
Naive Bayes	31.3
Logistic Reg.	47.5
~~*SVM*~~	---
Random Forest	56.4
XGBoost	**57.5**

5-fold cross-validation on 80% of data

Workflow - Tuning / Test

Model Tuning

[Model: XGBoost]

Goal: Maximize accuracy

Grid Search, via AWS:

- Max Depth
- Subsamples
- Child Weight
- Features per tree

Actual \ Predicted	Approved	Denied	Settled
Approved	20.3%	9.4%	1.6%
Denied	10.3%	34.6%	1.9%
Settled	5.0%	14.2%	2.8%

Accuracy: 57.6%

	Precision	Recall	F1
Approved	57.0%	64.9%	60.7%
Denied	59.4%	74.0%	65.9%
Settled	44.4%	12.6%	19.6%
Average	55.4%	57.6%	54.1%

Test on 20% Holdout

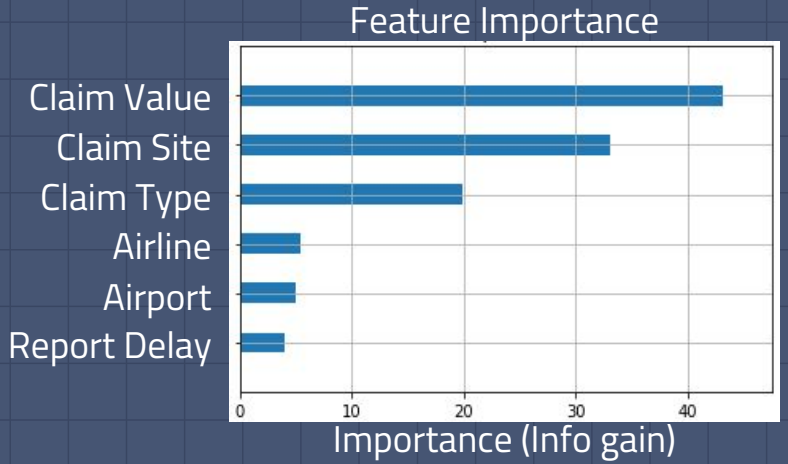
How do we use it?

Passenger / Insurance

- Predict your claim result
- Compare opportunity cost vs. expected return
- Adjust loan / benefits based on TSA judgment

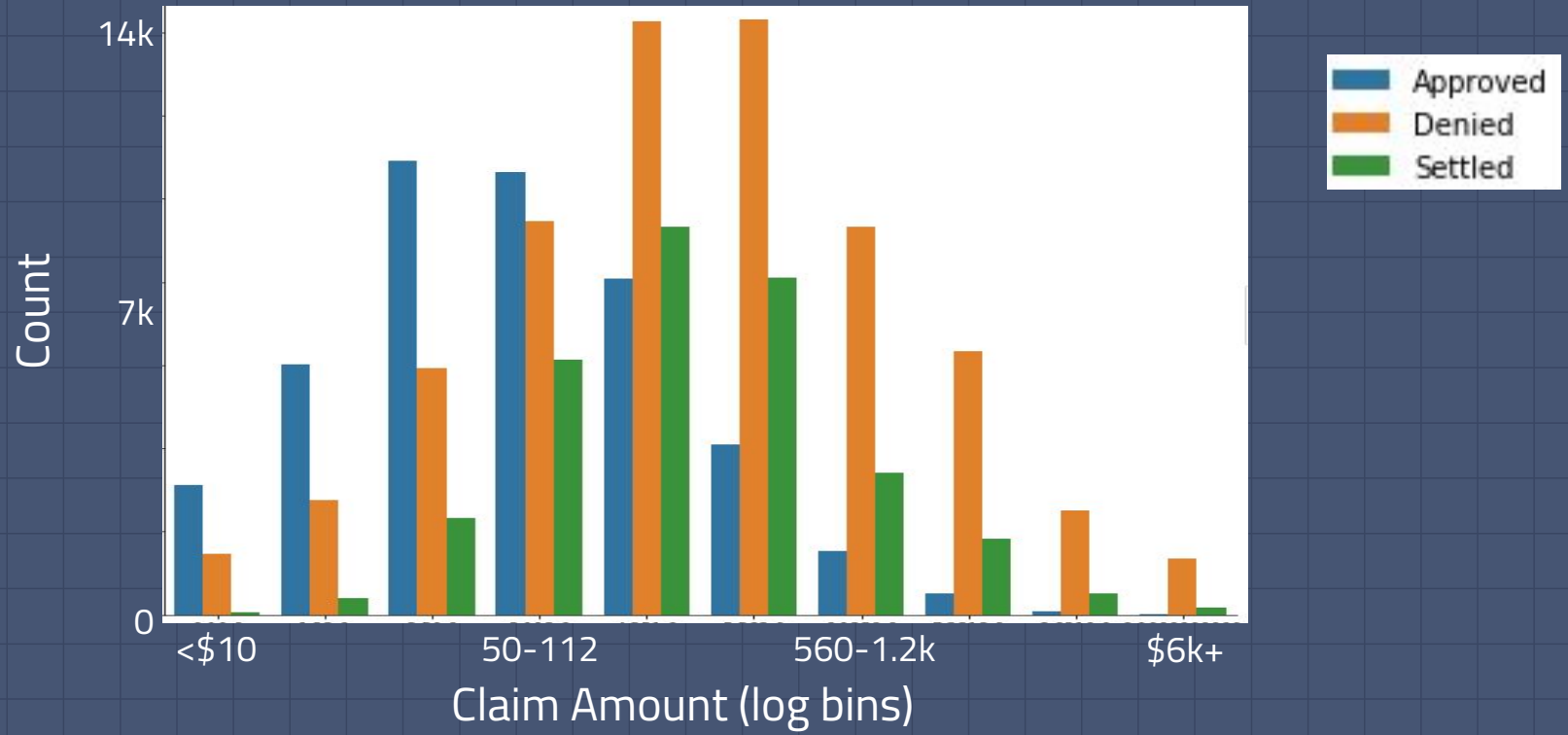
TSA

- Triage by predicted status and confidence
- Prioritize predictive features for mitigation efforts



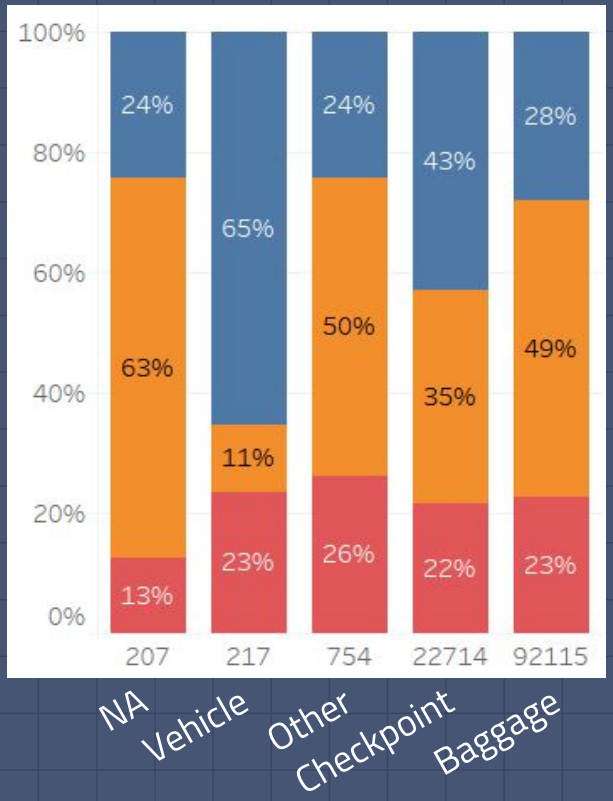
Predictive Features

Claim Amount

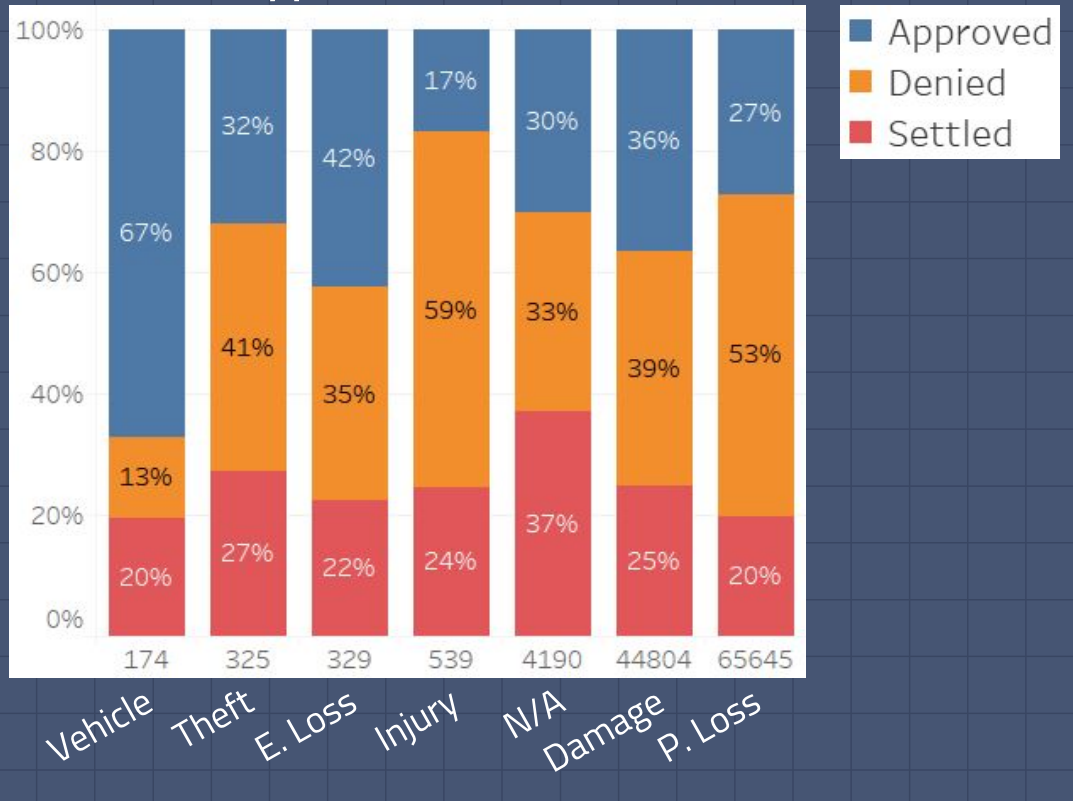


Predictive Features

Claim Site



Claim Type



Future Work

Add feature using Item Description

Optimize for performance on specific target class

- Pick a use case!
- Improve Approve / Deny / Settle depending on use case
- Resample or combine Settled cases with Approved

Host on interwebz

Thank you!