Yacht Insurance Claims Analysis and Prediction

CARLY CHAMBERLAIN SHARMA
SHARMA.CARLY@GMAIL.COM

WHAT WE'RE PREDICTING:

Will a yacht insurance policy have at least one claim during a 5 year period?

CONTENTS:

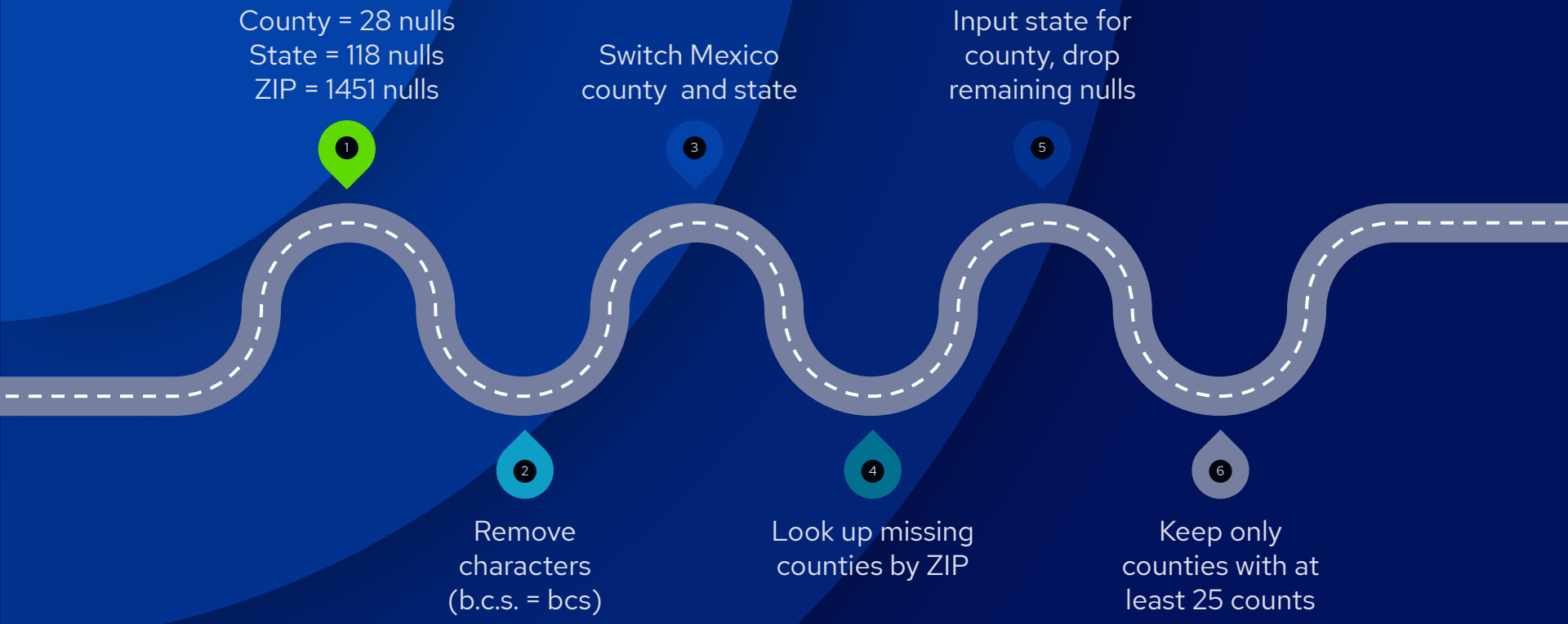
1. Data Cleaning

Limited values for target variable, had to keep as much data as possible

Using real world data: errors encountered

◎ Lower/uppercase disparities san francisco = San Francisco	◎ Different ways of expressing null values '-' and '.' and ' ' = NaN
◎ Incorrect values (human error or computer issue) DOB = 03/27/9062 ???	◎ Missing or "swapped" geographic data 'Mexico' = Mooring_County 'Mexico' = Mooring_State

Example: Editing Mooring Location



Wide range of values for categorical data:

- What to do with categorical data with a lot of unique values (builder, occupation, county)

OCCUPATION COLUMN

‘owns med device company’

‘business owner/president’

‘entrepreneur’

‘business owner’

2. Model Preprocessing

Feature engineering:

- Create an Age, Policy Length, and # of claims column
- Create columns for our target variables (multiclass and binary)
- Turn categorical columns into numerical (dummify)

3. Modeling

Multiclass classification attempt:

- Wanted to predict whether a policy would have 1, 2, or 3 claims.
- Too little data

0 Claims = 5836 (92.0%)
1 Claim = 421 (6.6%)
2 Claims = 68 (1.0%)
3 Claims = 15 (0.2%)

Binary Classification Steps

RUN INITIAL MODELS → OVER UNDER SAMPLE → FEATURE SELECTION → FINE TUNE BEST MODEL

0 Claims = 5836 (92.1%)

At least 1 Claim = 504 (7.9%)

Metrics Used:

- **Accuracy**: how well the model predicts either a claim or no claim (skewed to majority class)
- **Weighted F1 Score**: Weighted avg. of precision and recall.
- **Recall: Calculates how many of the actual positives a model captures through labeling it as positive.**

STEP 1: Run initial models

- KNN, Logistic Regression, Random Forest, Extra Trees
- Ran the four models with basic or default parameters to get baseline scores

STEP 2: Over and Under Sample

- RandomOverSampler
- SMOTE OverSampler + RandomUnderSampler

BASELINE COUNT:
0: 5836
1: 504

ROS COUNT:
0: 4669
1: 933

SMOTE + RUS COUNT:
0: 932
1: 466

Initial Results:

	Accuracy	Recall	Weighted F1 Score
KNN(plain)	0.913249	0.128713	0.893392
LR(plain)	0.919558	0	0.881779
RF(plain)	0.936909	0.227723	0.918876
ET(plain)	0.925868	0.257426	0.912519
KNN(oversample)	0.867508	0.336634	0.876073
LR(oversample)	0.896688	0.059406	0.876649
RF(oversample)	0.932965	0.257426	0.917971
ET(oversample)	0.926656	0.237624	0.911725
KNN(SMOTE/Under)	0.765773	0.366337	0.809979
LR(SMOTE/Under)	0.806782	0.277228	0.834278
RF(LR(SMOTE/Under))	0.895899	0.376238	0.897266
ET(LR(SMOTE/Under))	0.861987	0.39604	0.874726
LRCV(SMOTE/Under)	0.827287	0.267327	0.847037

STEP 3: Further feature selection

- Use only the top 30 most important features
- Drop ‘policy_length’ and ‘New/Renl/Endt/Can/Flat’

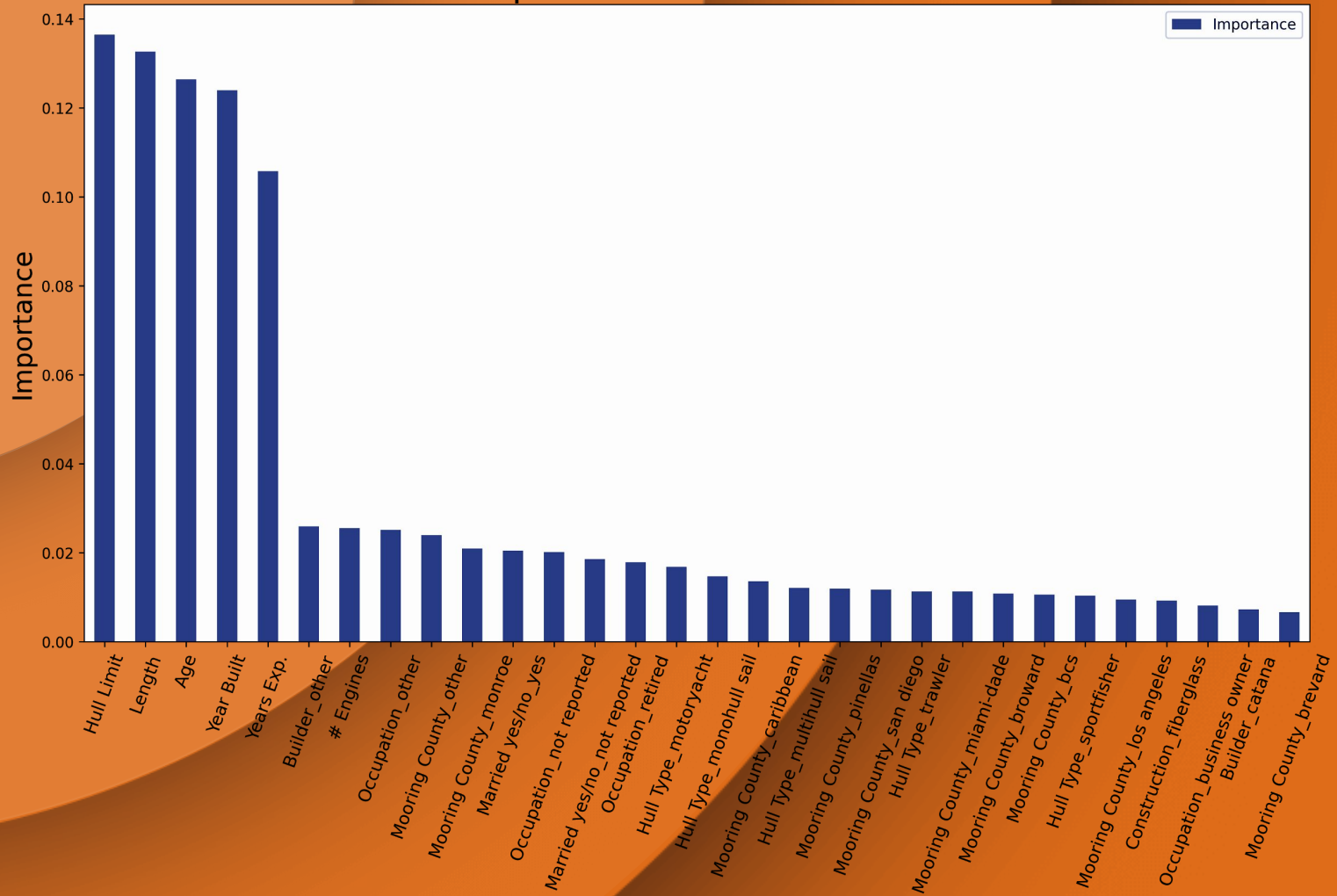
4. Results

STEP 4: Fine tune best model

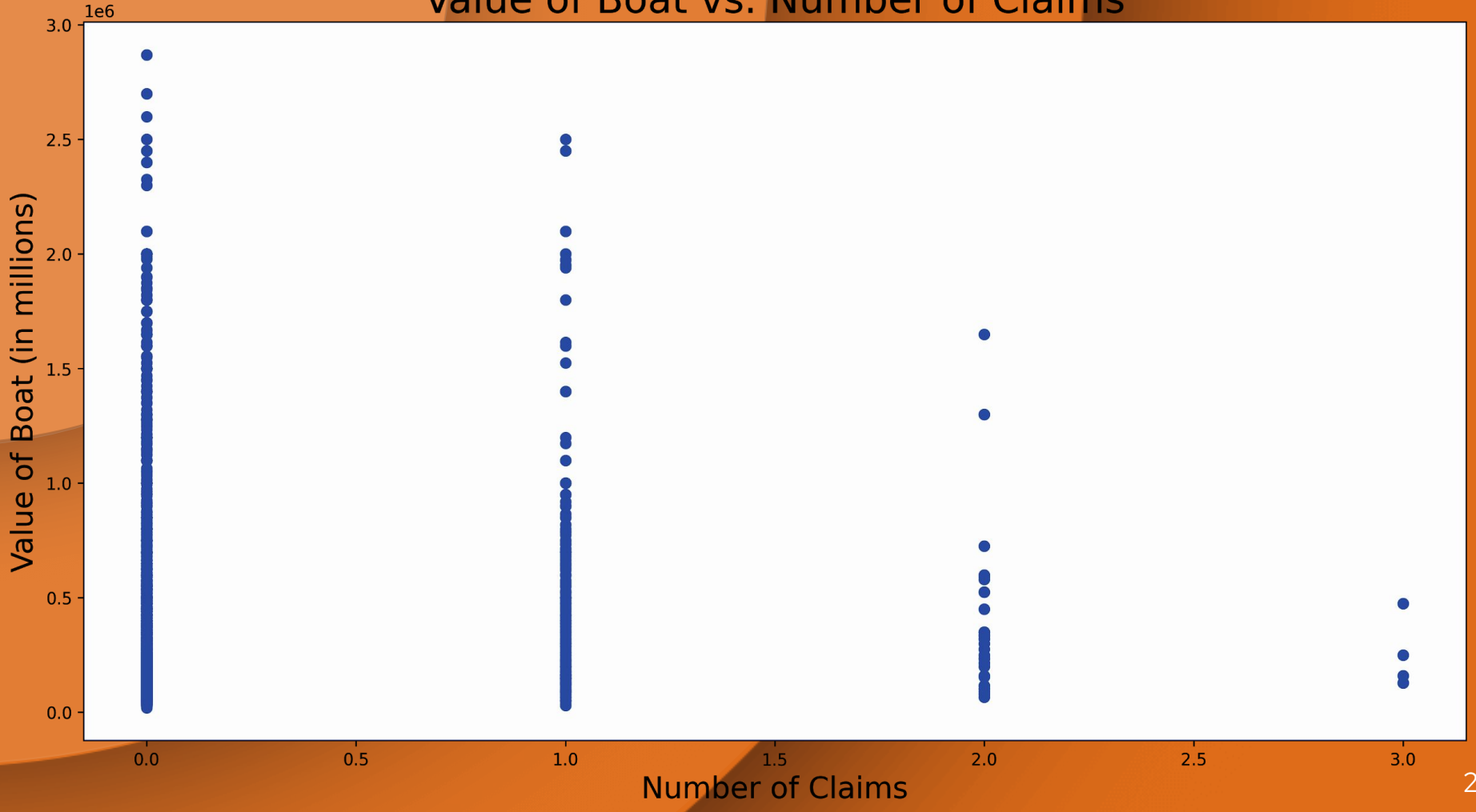
- Gridsearch over each model
- Ultimately chose Random Forest

	Accuracy	Recall
RF(Training)	1	1
RF(Testing)	0.878549	0.445545
RF_Feature_Drop(Training)	0.962089	0.924893
RF_Feature_Drop(Testing)	0.865142	0.445545
RF_Top_30(Training)	0.959227	0.924893
RF_Top_30(Testing)	0.855678	0.465347

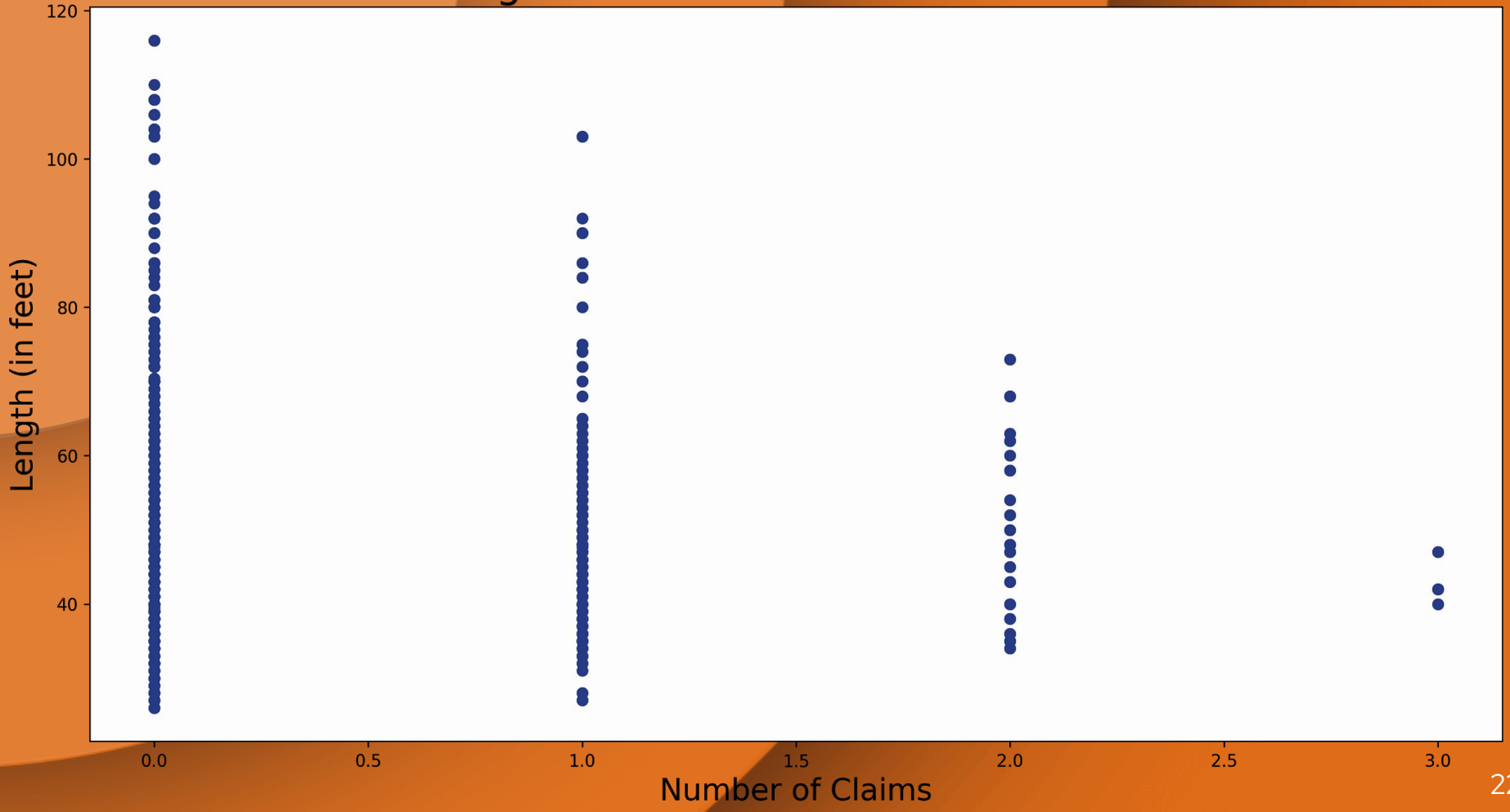
Most Important Features of RF Model



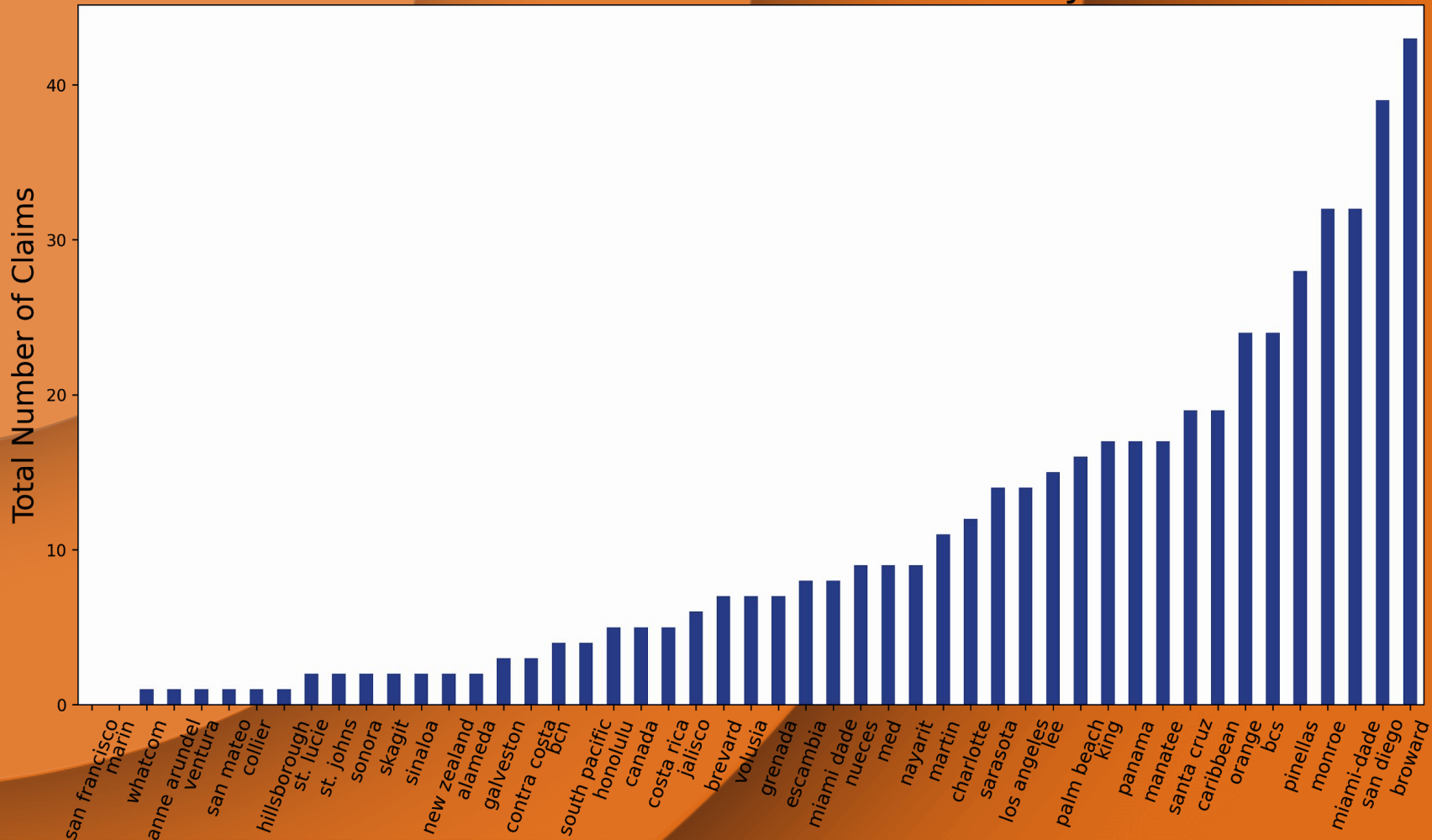
Value of Boat vs. Number of Claims



Length of Boat vs. Number of Claims



Total Number of Claims Per County



Conclusions and Recommendations:

- Research specific claim cases with regards to the value, length, age of the boat and age/years of experience of owner.
- Charge more in premiums for hurricane prone areas (Florida, Caribbean).
- Overhaul your data systems.

Further research/analysis

1. Redo models after working with client to: Categorize Occupation, Builder, and Counties
2. With more data I'd like to break down types of claims (theft, operational, weather). Could be interesting to view in respect to the features.
3. Look at pandemic stats. Are boats being used less?
4. Auto-generate premium costs based on likelihood of a claim.

Thank you. Questions?