TRAVEL INSURANCE CLAIM PREDICTION

Abeselom Fanta

February 22, 2022

OUTLINE

- OVERVIEW
- BUSINESS AND DATA UNDERSTANDING
- MODELING
- RESULTS
- CONCLUSION
- NEXT STEPS

OVERVIEW

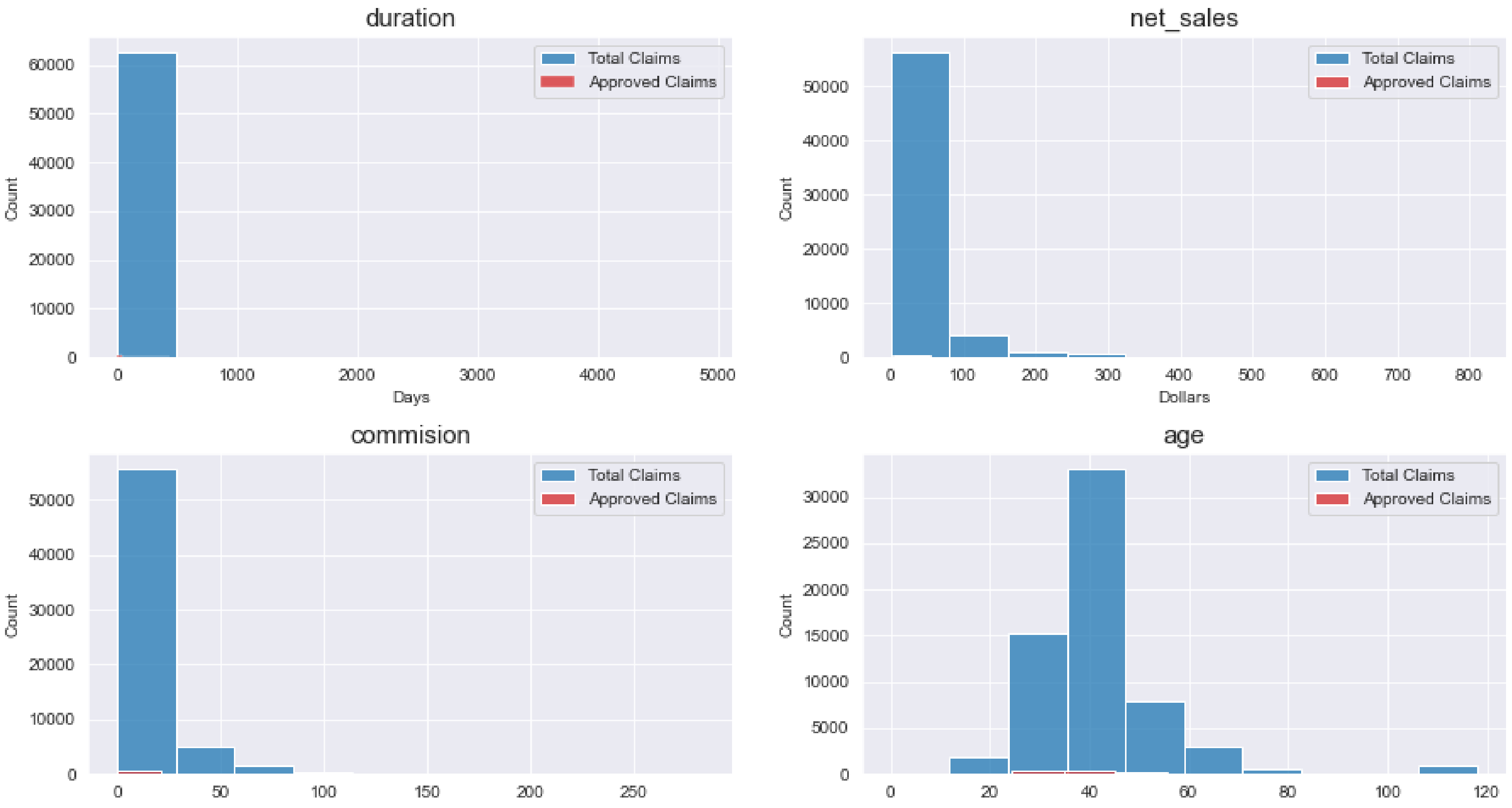
- Travel insurance data from a third-party company in Singapore
- Claim approval depends on ten independent features
- Predictions made based on:
  - Age, duration and destination requested of the insuree
  - Commission and net sales of insurance policy
  - Insurer agency, agency type and distribution channel
  - Product name (insurance type)

BUSINESS AND DATA UNDERSTANDING

- Data collected from Kaggle
- This project will try to address the following questions:
  - What is the nature of the data?
  - Are additional steps needed to reduce data size for processing?
  - Which machine learning model provides the highest true positive rate?
  - Which features are important for prediction?
  - Can a web deployable model be developed?

DATA UNDERSTANDING

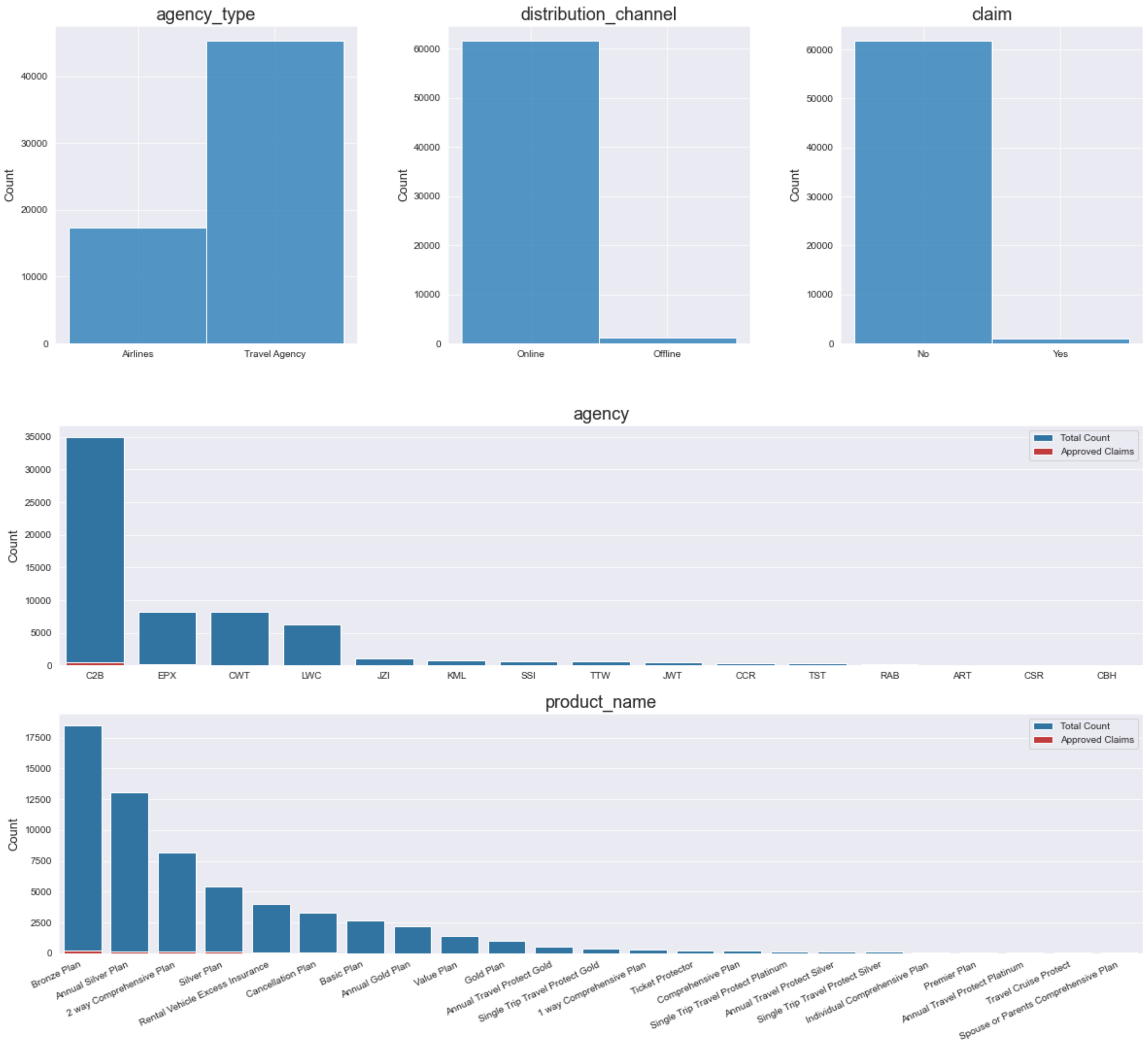
Distribution for numerical features



DATA UNDERSTANDING

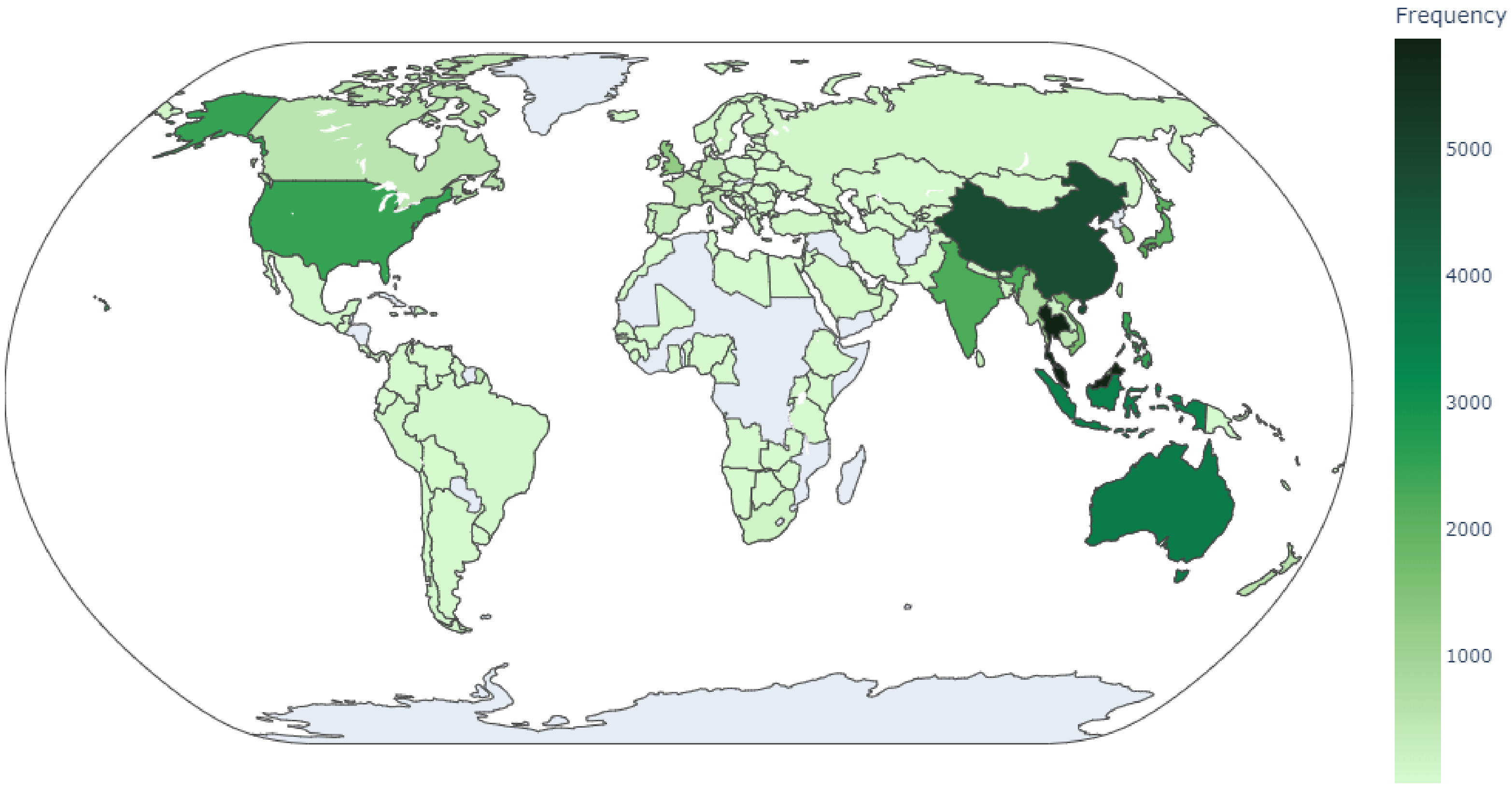
Distribution for categorical features

Data is highly imbalanced



DATA UNDERSTANDING

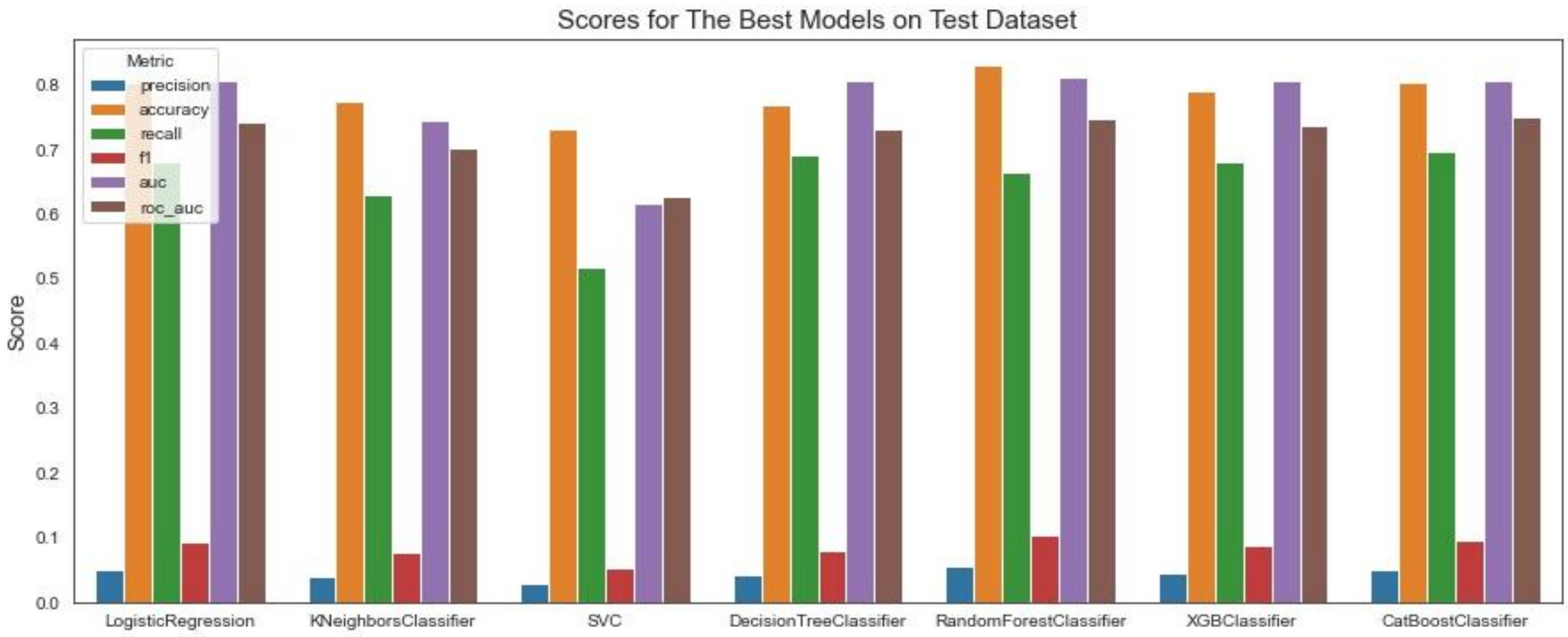
Claims based on destination
(excluding Singapore)



MODELING

- Destinations replaced by continents
- Seven machine learning models used (logistic regression, K nearest neighbors, support vector machine, decision tree, random forest, gradient boost models: XGBoost and CatBoost)
- A total of 4,200 models trained (pretrained model available for download)
- Models evaluated for accuracy, recall and true positive rate
- Best models chosen from the machine learning models used
- Final model selected
- Variations of final model with important features and destinations compared
- The best performing model saved to file for the web app

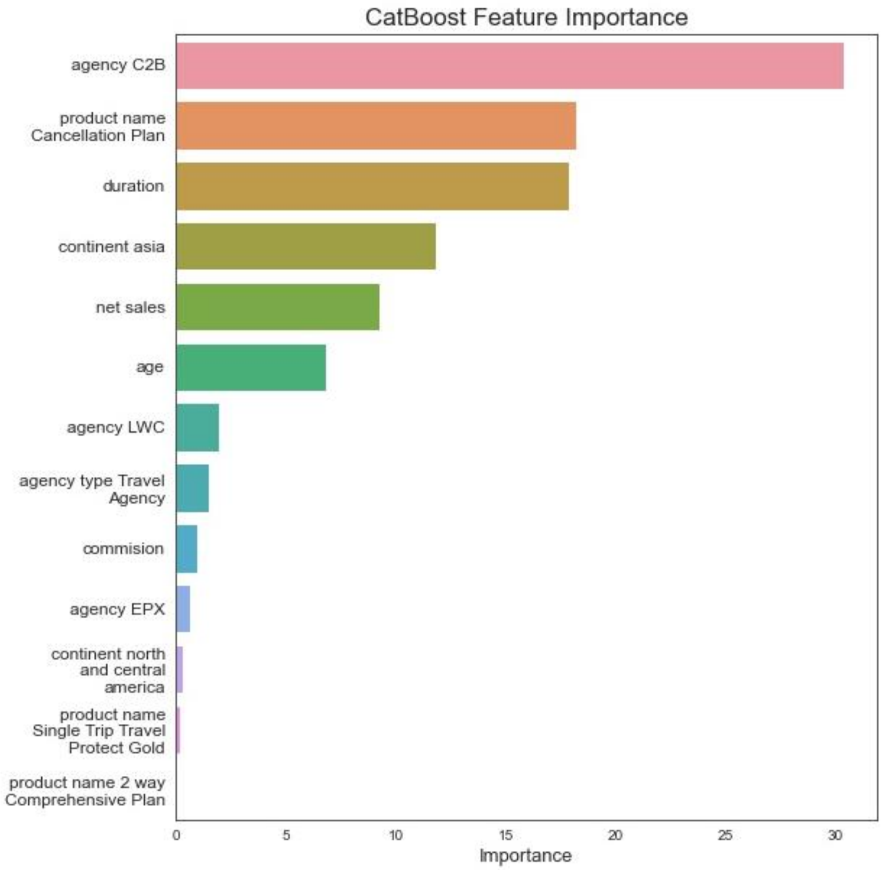
RESULTS

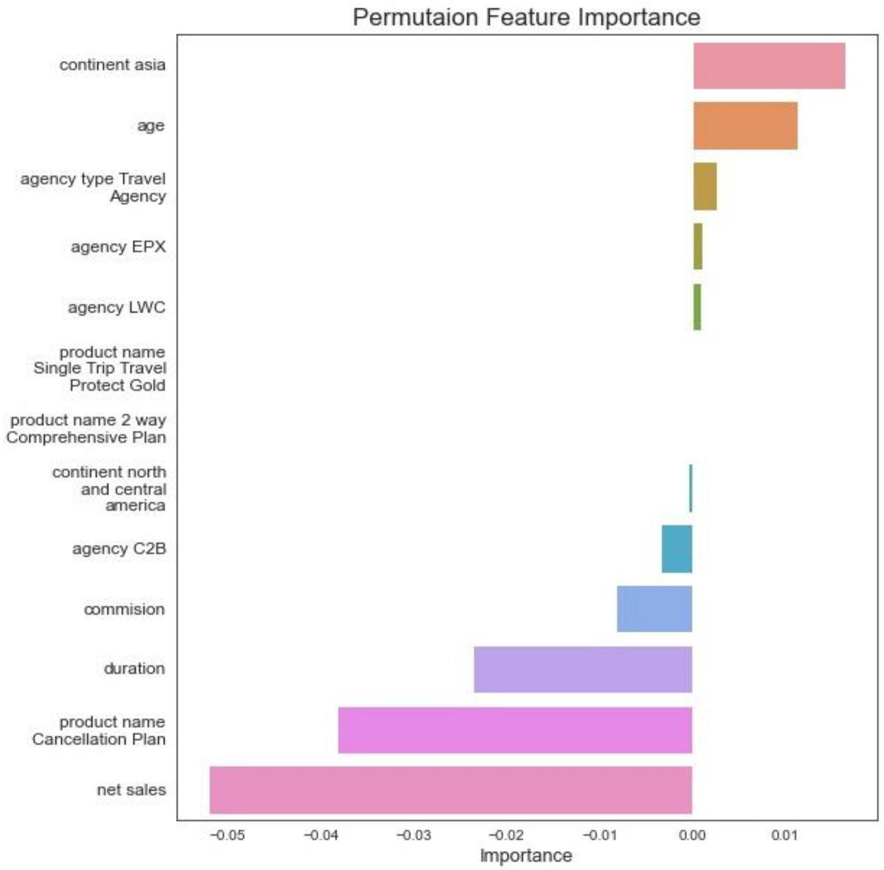


MODEL COMPARISON

Random forest and CatBoost showed the highest metrics

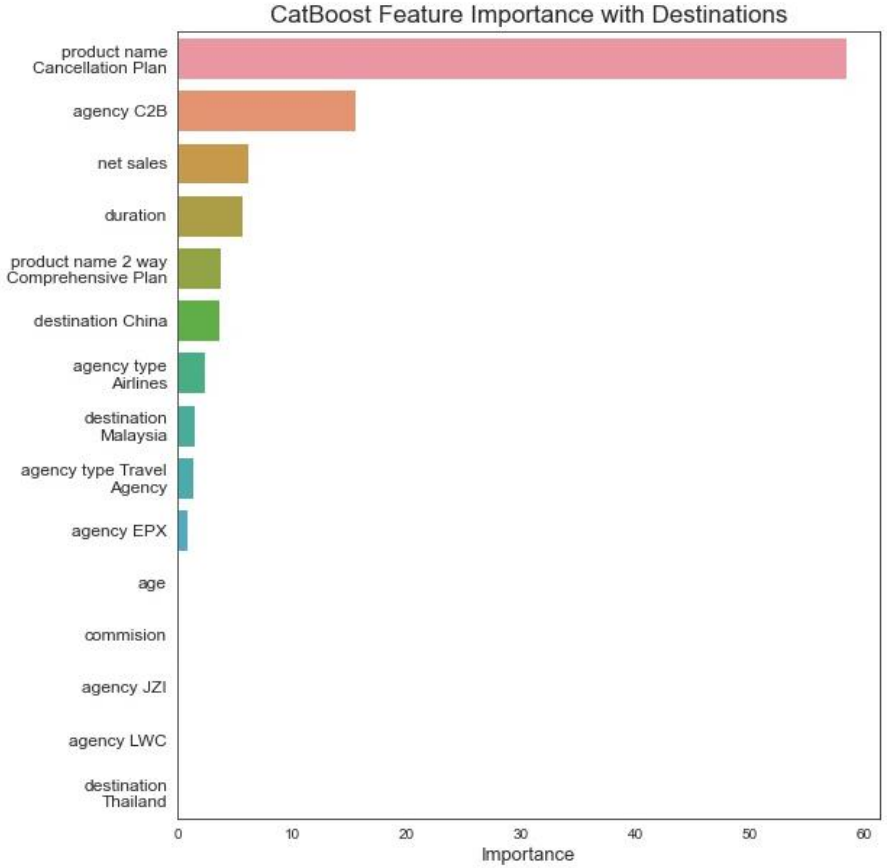
CatBoost selected as a final model

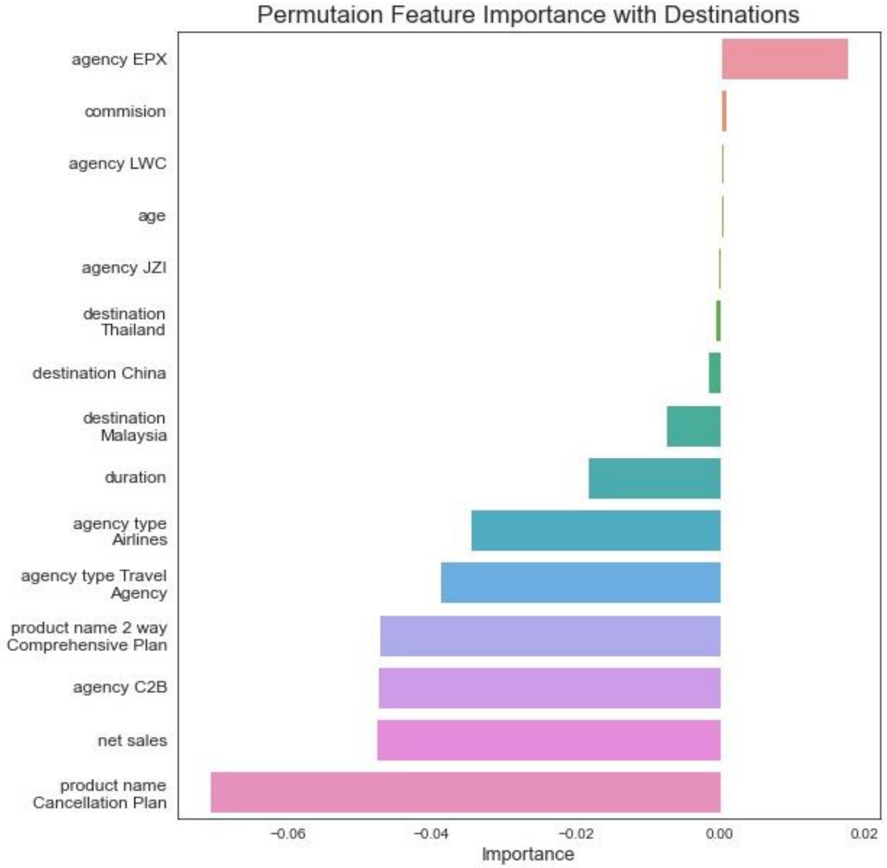




FEATURE IMPORTANCE

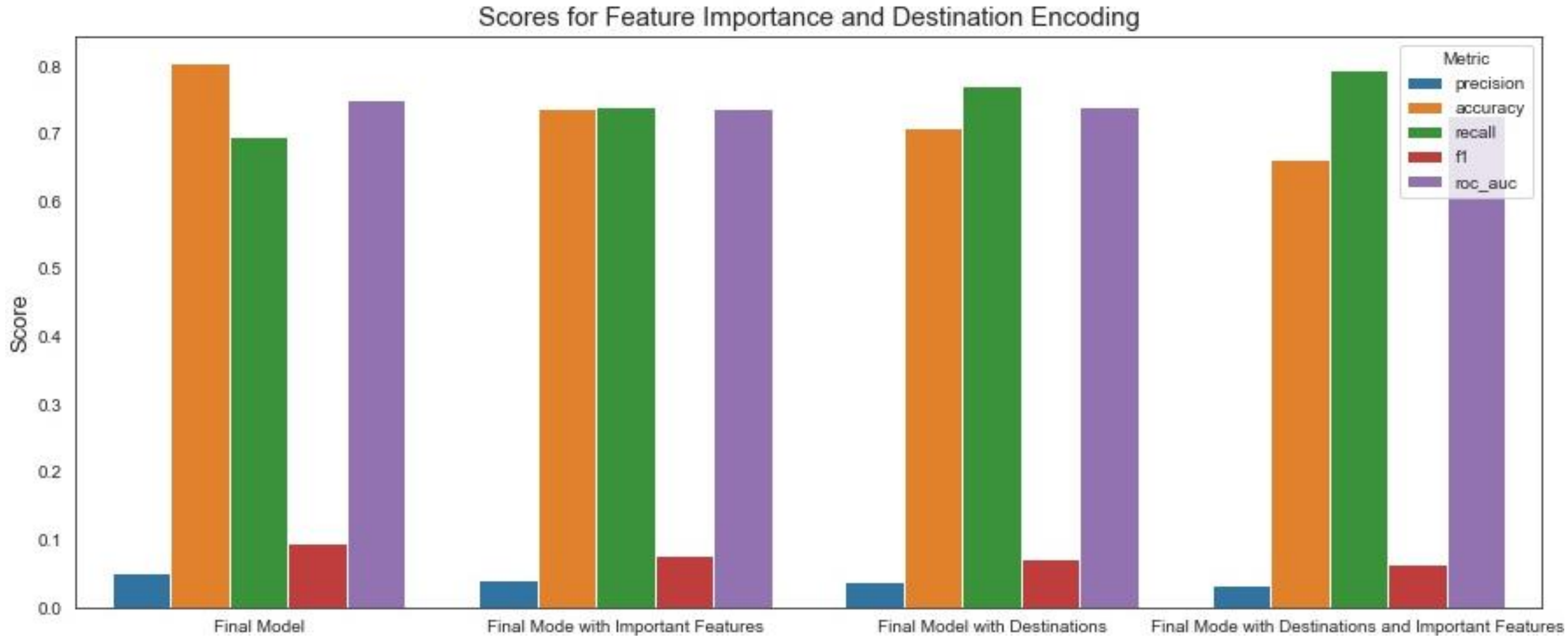
Some features had stronger influence over the model prediction





FEATURE IMPORTANCE WITH DESTINATIONS

Adding destinations as features caused importance changes but most features remained unchanged

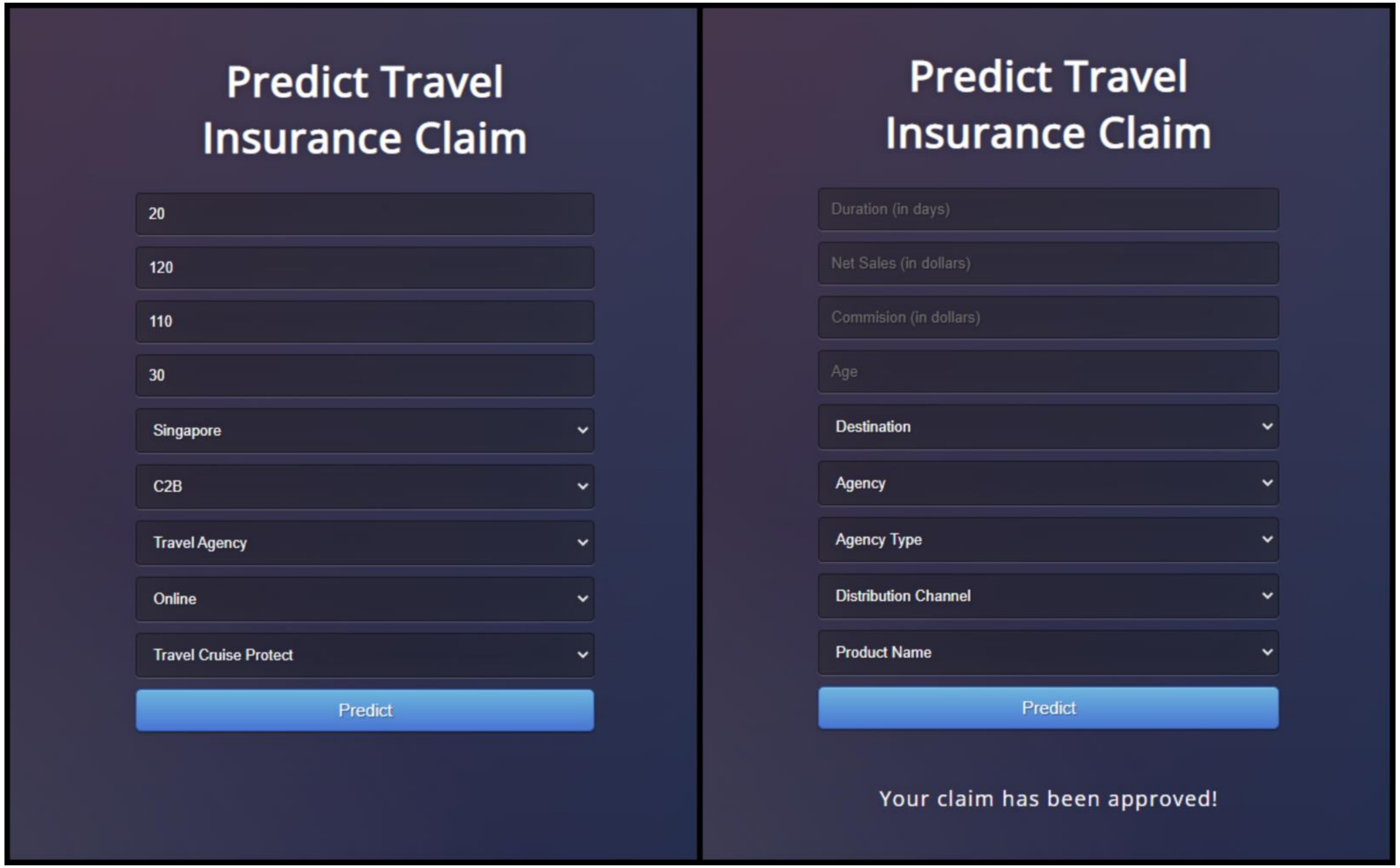


FINAL MODEL VARIATIONS

Final CatBoost model with no additional features performed the best

WEB DEPLOYMENT

Final CatBoost model saved for a web-based claim predictor app



CONCLUSION

- Using random forest and CatBoost yielded the highest true positive rates
- Very few features appeared to be more important
- Tradeoff between recall and true positive reduces accuracy
- Using destinations instead of continents reduces metrics

NEXT STEPS

- Gather mode data with more diverse destinations
- Balanced claim approval to denial rate
- Complete gender of insuree
- Date of insurance claim request
- Houses with excellent views and condition cost more than the rest.

THANK YOU

Abeselom Fanta

GitHub: @afanta-fi