Exploring the feasibility of imputation techniques for the Commodity Flow Survey (CFS)

2024 Federal Committee on Statistical Methodology

October 24, 2024
College Park Marriott Hotel & Conference Center
3501 University Blvd E, Hyattsville, MD 20783

Gritiya Tanner

United States® Census Bureau

Disclaimers

Any opinion and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Data Management System number: P-7504831, Disclosure Review Board (DRB) approval numbers: CBDRB-FY24-ESMD002-018.

Commodity Flow Survey (CFS)

- Data on the movement of goods within the United States
- Every 5 years as part of the Economic Census (mandatory)
- Joint effort by U.S. Census Bureau and the Bureau of Transportation Statistics (BTS)
- Main input to BTS's Freight Analysis Framework

Questionnaire

Shipment characteristics

- Value ($)
- Weight (pounds)
- Type of commodity (description)
- Temperature controlled (Y/N)
- Hazardous material number (UN/NA)
- **Domestic mode of transportation**
- Export mode of transportation
- Domestic destination (state, city, zip)
- Export destination (country, city, postal code)

Mode of transportation codes

1 - Parcel, U.S.P.S, or courier

2 - Company-owned truck

3 - For-hire truck

4 - Railroad

5 - Inland water

L - Great Lakes

6 - Deep sea

7 – Pipeline

8 - Air

9 - Other mode*

0 - Unknown*

C - Customer pick-up

*Please specify domestic mode of transportation used

Note: For customer pick-up, use the customer's address on file or use the origin ZIP Code if the final destination ZIP Code is unknown.

Research objective

- Redesign of 2022 data collection
  - 2017 – respondent ask to provide a sample of shipments for 4 assigned weeks
  - 2022 – respondents were given a choice
    - provide a sample of shipments for 4 assigned weeks
    - provide ALL shipments from for 4 assigned weeks

- In 2022, 16 times more shipments were collected compared to the previous cycle in 2017.

- Rule-based imputation approach is no longer sufficient. It can be used to impute only 3% out of 10% missing or unknown mode of transportation.

- Which machine learning classifier model best predicts domestic mode of transportation?

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Detour – Machine Learning Basics

Accuracy: Overall percent that the model predict target class correctly.

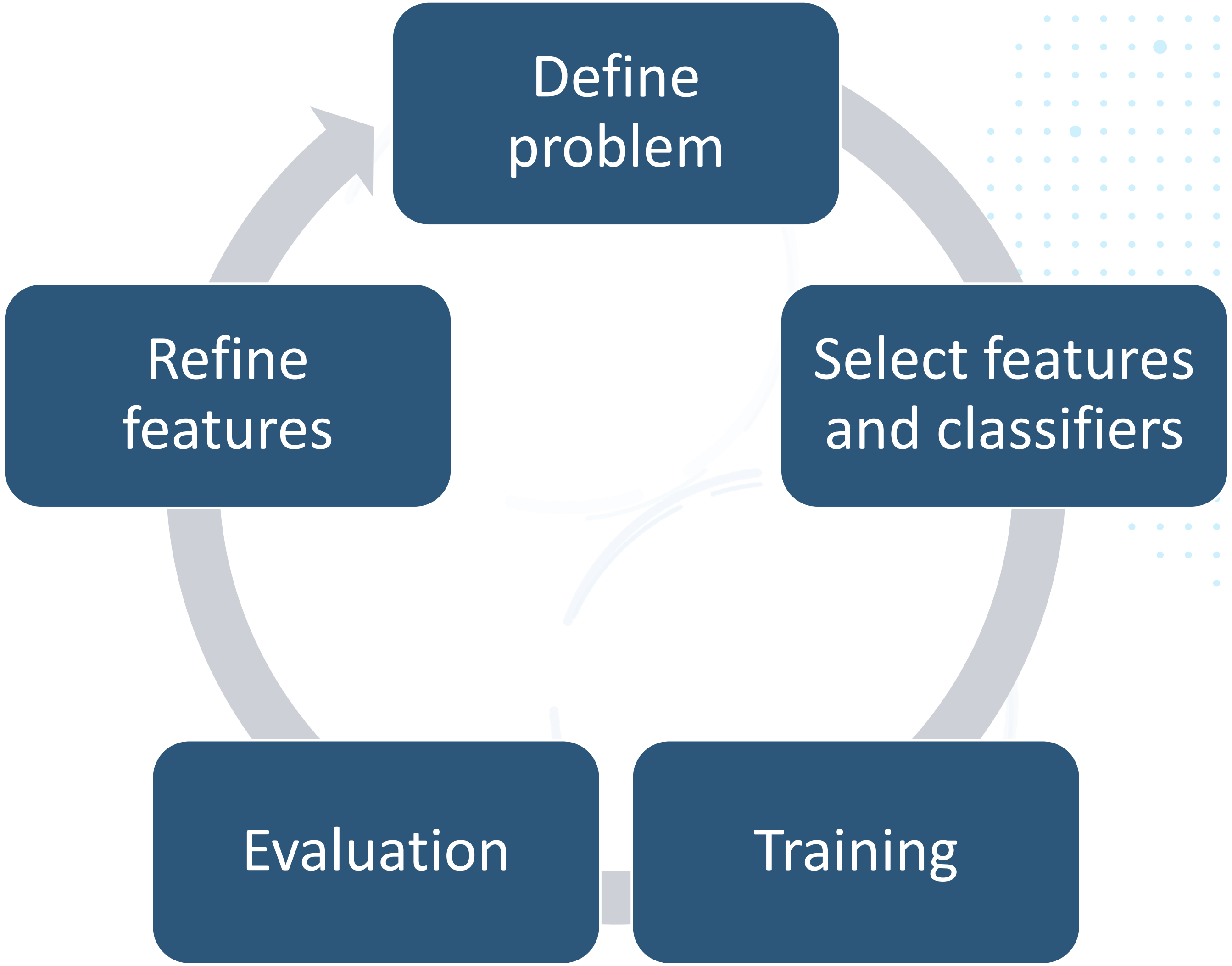
Classification report:

- Precision: Percent that the model predicts target class correctly.
- Recall: Percent that the model find or catch the correct target class.
- F-1 Score: Percent of a harmonic mean of precision and recall

  F1 Score = 2*[(Recall * Precision) / (Recall + Precision)]

Confusion Matrix: Comparing predicted value against actual value.

Cycle of machine learning tasks



Define problem

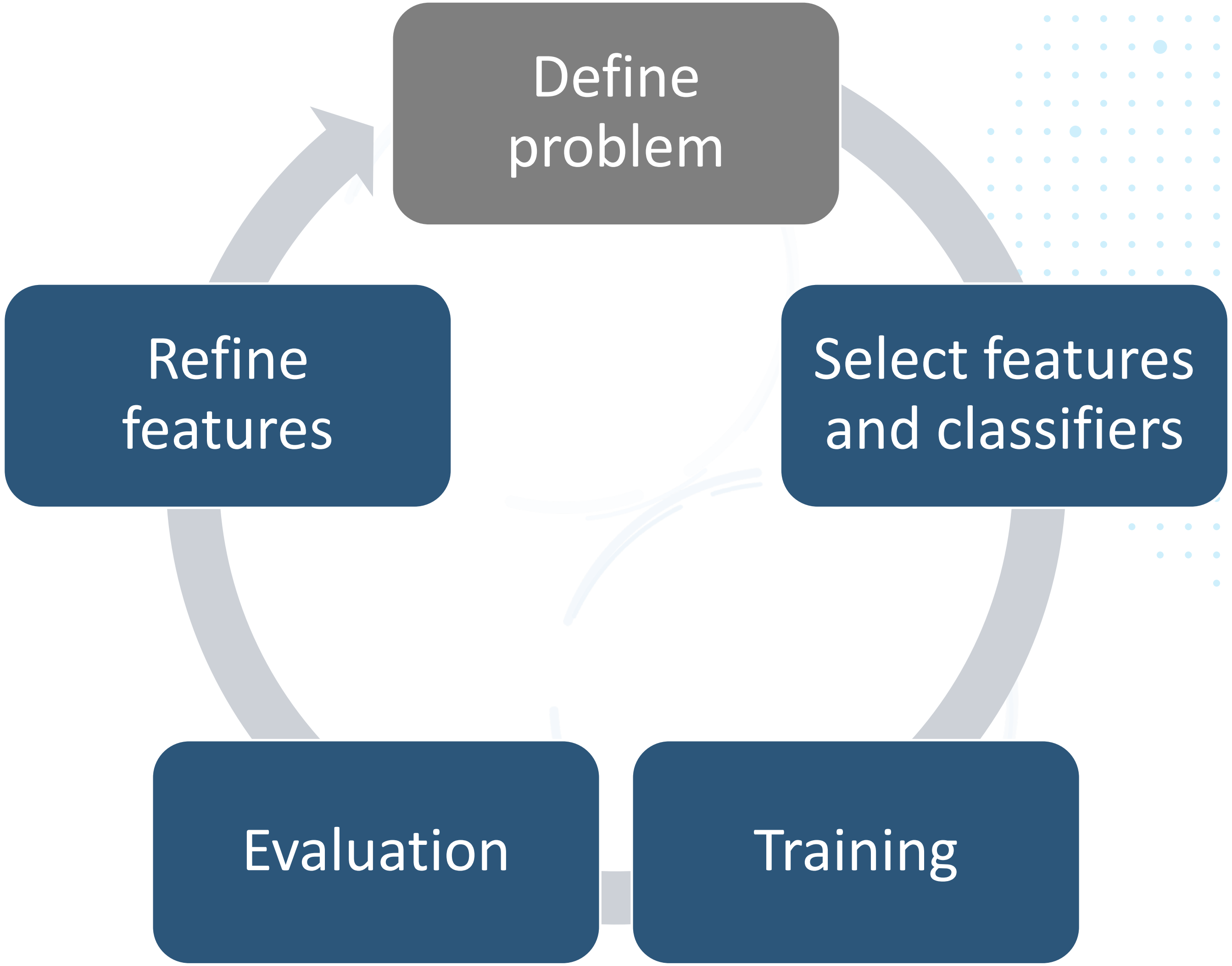
Select features and classifiers

Training

Evaluation

Refine features

Cycle of machine learning tasks



Define problem

Select features and classifiers

Training

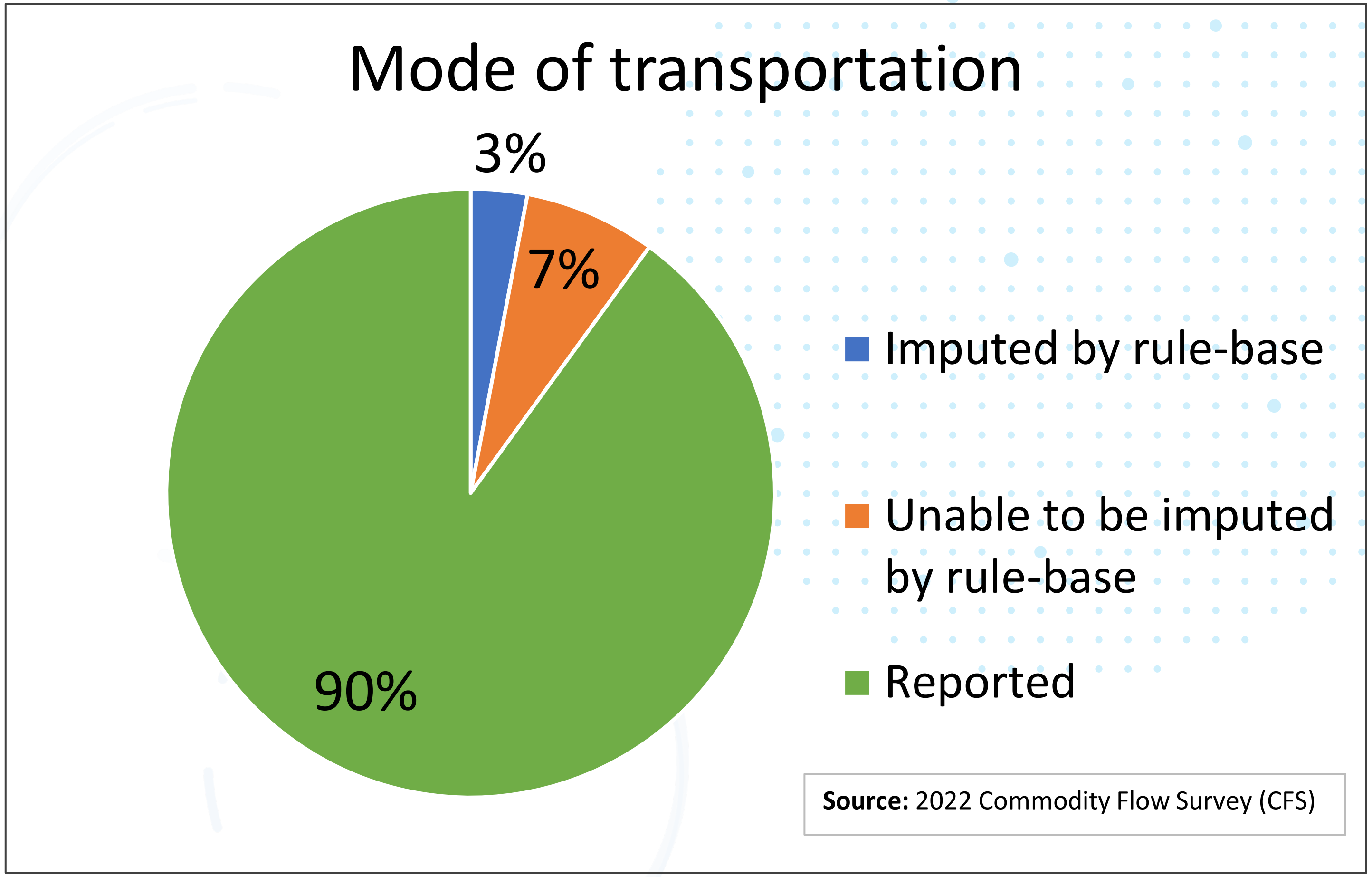
Evaluation

Refine features

Define problem

We collect about 100 million shipments in 2022 CFS.

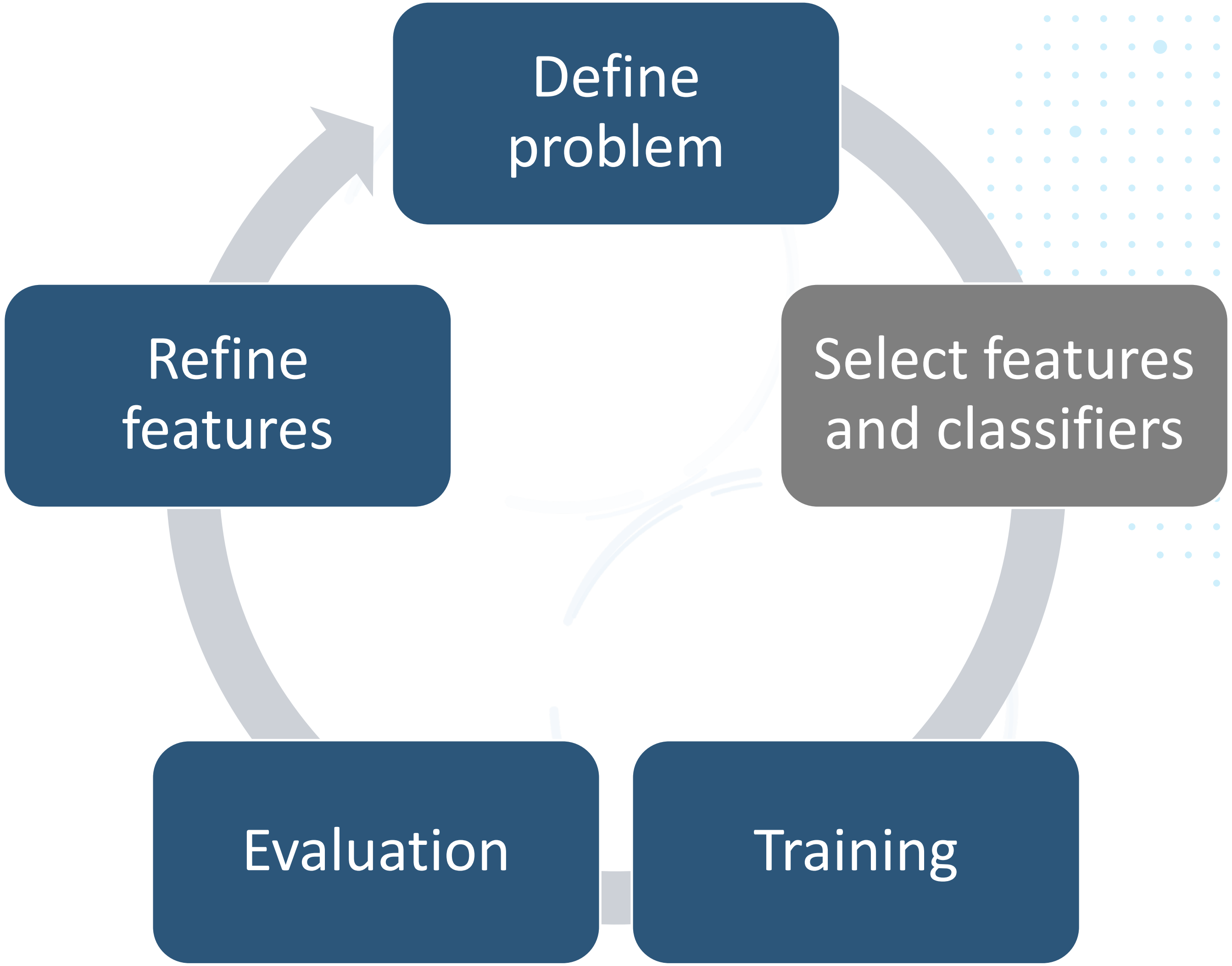
- 10% missing mode or report unknown mode
- 3% imputed by rule-base approach
- 7% unable to be imputed by rule-base
- Explore machine learning to impute domestic mode of transportation on 7%.



Source: 2022 Commodity Flow Survey (CFS)

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Cycle of machine learning tasks



Define problem

Select features and classifiers

Training

Evaluation

Refine features

Select initial features

Features

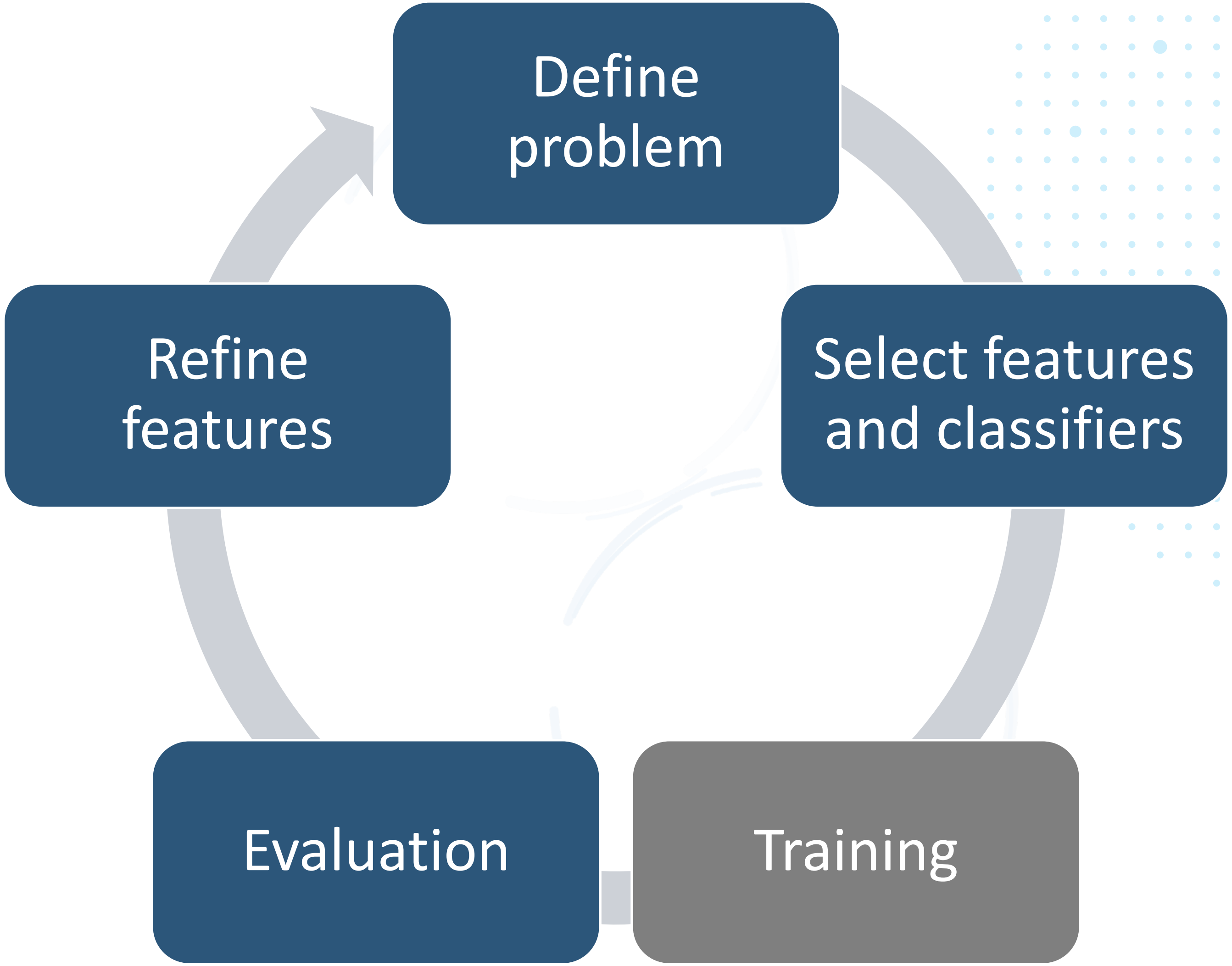
- Categorical features:
  - Commodity
  - Origin Zip
- Continuous features:
  - Distance
  - Value
  - Weight

Select appropriate classifiers

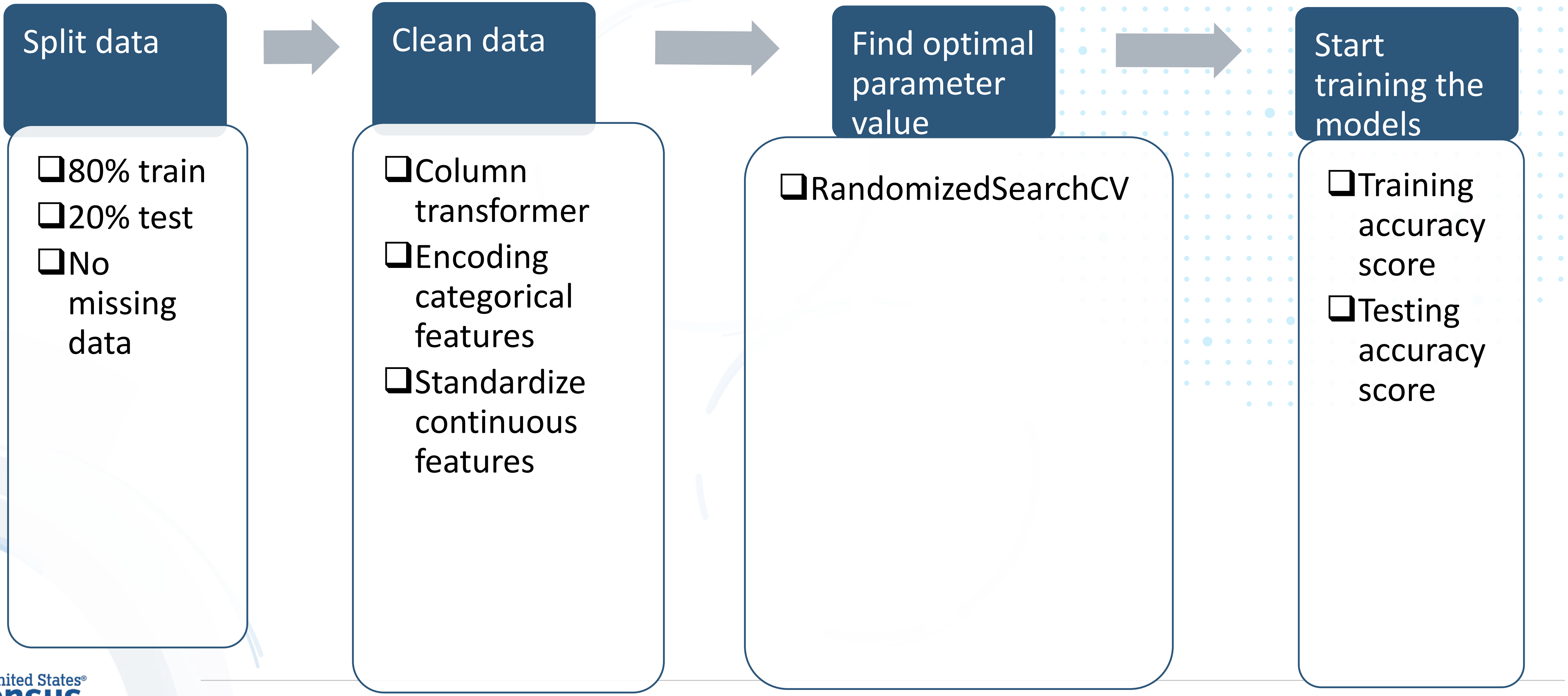
Supervised learning classifier

- ❑ Dummy
- ❑ Decision tree
- ❑ K-nearest neighbor
- ❑ Naive bayes
- ❑ Support Vector Machine

Cycle of machine learning tasks



Training classifier models



Split data

- □80% train
- □20% test
- □No missing data

Clean data

- □Column transformer
- □Encoding categorical features
- □Standardize continuous features

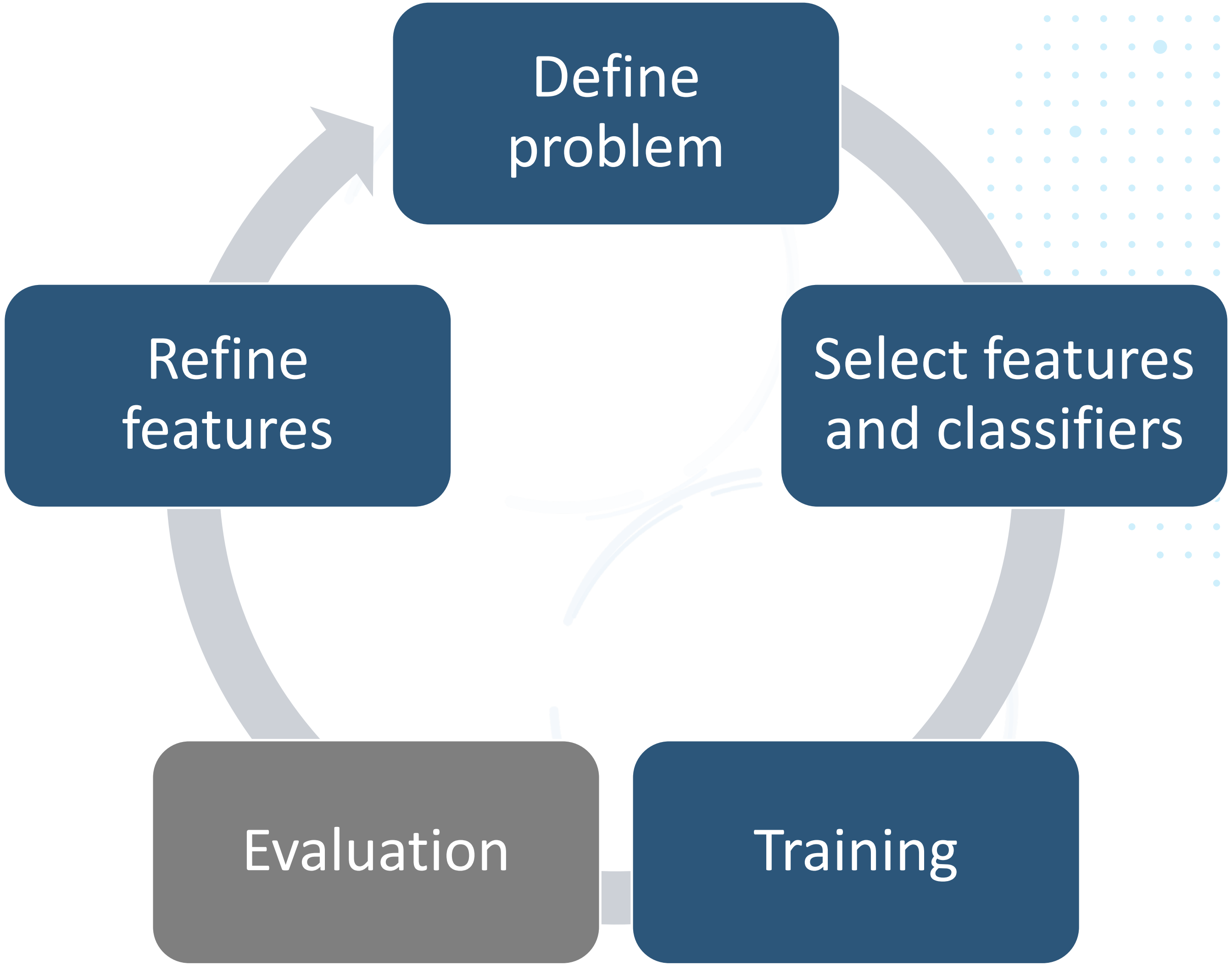
Find optimal parameter value

- □RandomizedSearchCV

Start training the models

- □Training accuracy score
- □Testing accuracy score

Cycle of machine learning tasks



Define problem

Select features and classifiers

Training

Evaluation

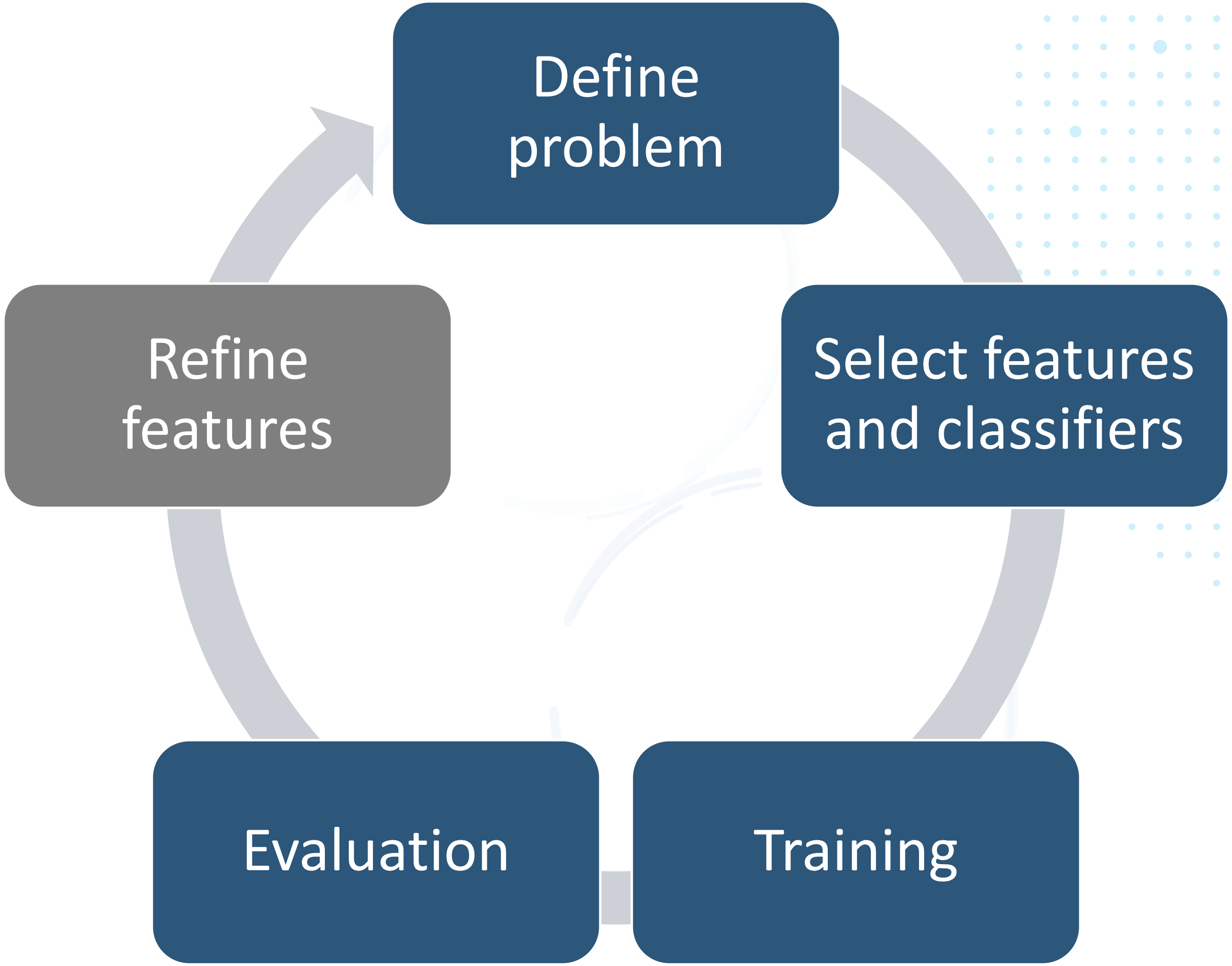
Refine features

Evaluating models using 5-Fold Cross Validation

Model	Average Accuracy score	Performance (Runtime on laptop using 10% sampling training data)	Features
Dummy classifier	54%	1 Minute	All features
Decision tree	**93%**	**7 Minutes**	**All features**
K-nearest neighbor	76%	5 Minutes	Only numeric features
Naïve bayes	55%	2 Minutes	Only numeric features
Support vector machine	n/a	Take over 3 hour	All features

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Cycle of machine learning tasks



Define problem

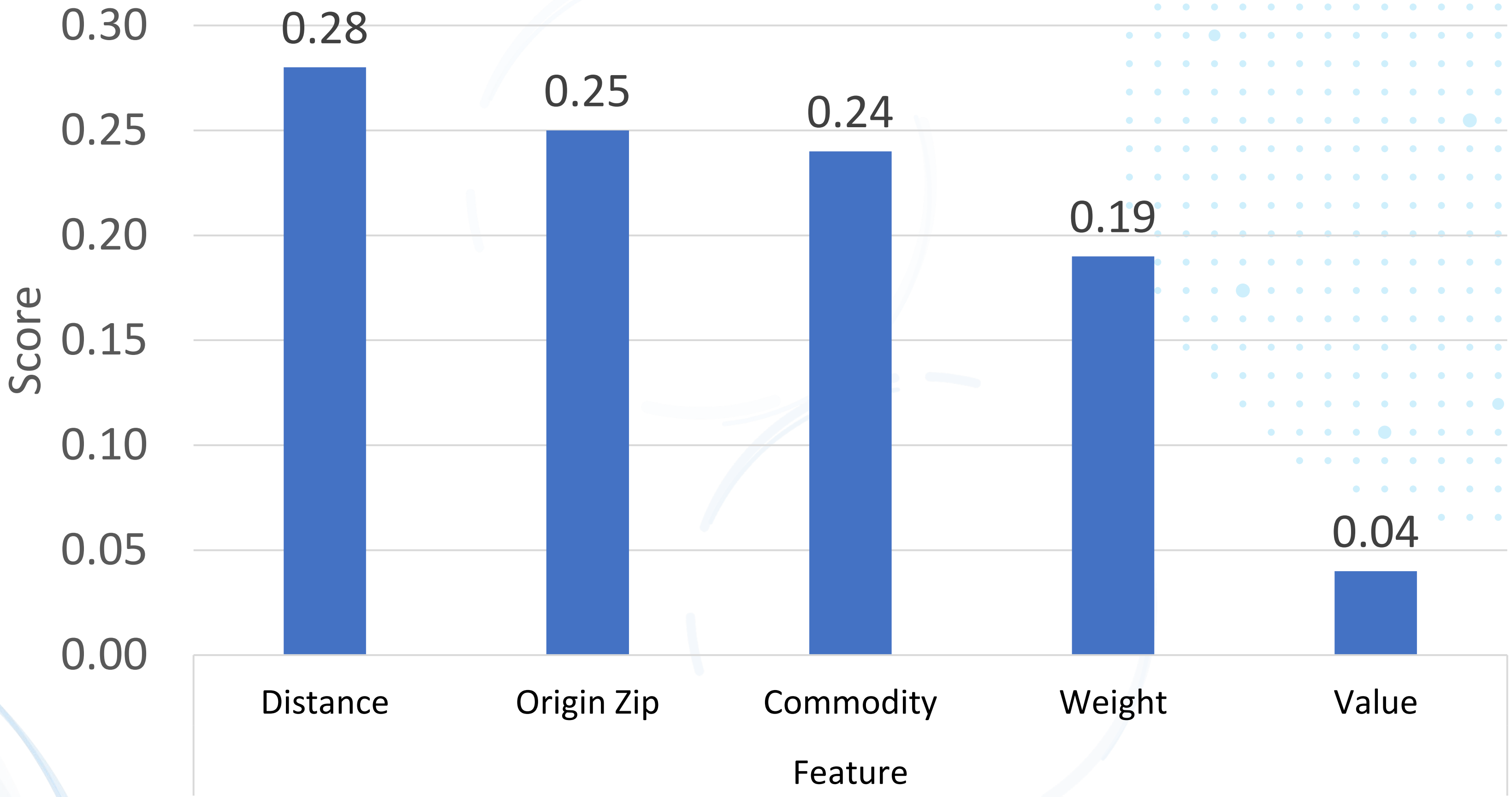
Select features and classifiers

Training

Evaluation

Refine features

Feature importance



Source: 2022 Commodity Flow Survey (CFS)

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Classification report

Using the F1-score to see how well the model works.

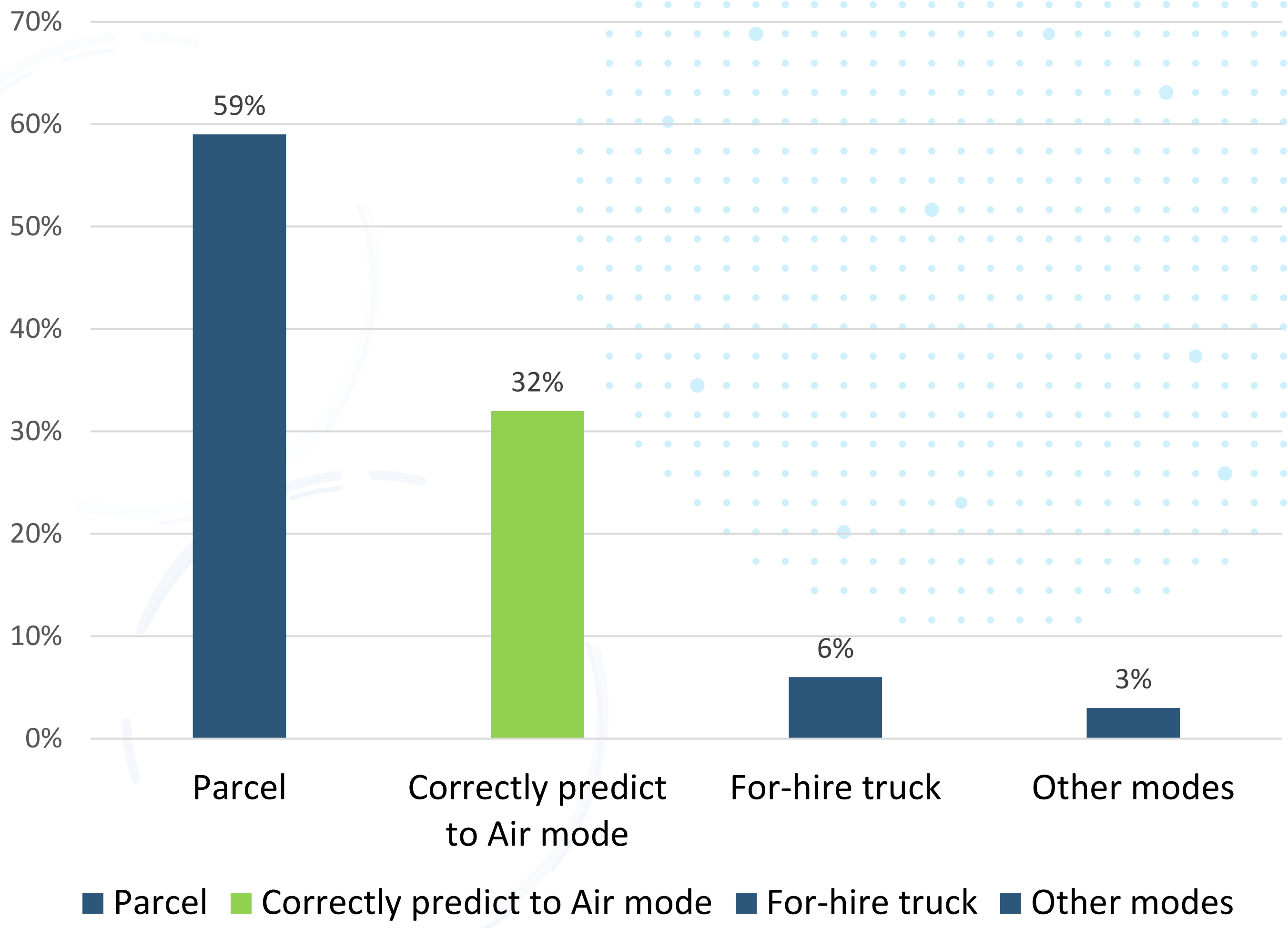
- The model performs well with the F1-score between 70% and over 90% on the following modes:
  - Parcel, U.S.P.S. or courier
  - Company owned truck
  - For-hire truck
  - Customer pick-up
- The model performs worst on the **air mode** with the F1-score 40%. (Precision 55% and Recall 32%)
- The air mode is one of our top modes, this prompts us to do further investigation using confusion matrix.

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Confusion matrix

Confusion matrix helps narrow down to where the model makes a mistake.

- 59% predict to **parcel** mode instead of air mode.
- 6% predict to for-hire truck mode.
- Could model predict toward dominant mode?
- Need investigate imbalance data issue.



Source: 2022 Commodity Flow Survey (CFS)

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Handling imbalanced data

- Review data distribution: The air mode represents up to only 0.65% of the shipments indicating to us that we have imbalanced data.
- Adjusted the model by adding a parameter, class_weight='balanced’, to account for imbalanced data in the model.
- The accuracy score of the new model dropped to 88% which is still very close to the initial model 93%.
- It indicated that the decision tree classifier is not effected by imbalanced data.
- Revisit the feature importance result to help improve the accuracy score of the minority modes and still maintain high overall accuracy score.

Summary

- **Decision tree classifier** is the best prediction model for domestic mode of transportation compared to other selected models.
  - High average accuracy score.
  - Acceptable runtime.
  - Does not require data preprocessing.
  - Handles imbalanced data well.
- Based on the initial selected features, the model works very well predicting most of the domestic mode of transportation data.
  - The model predicts 98% of the data with f1-score between 70% and over 90%.
- As a reminder that this research is still in progress, it is necessary to go through the cycle of machine learning tasks multiple times to continue to improve probability prediction of the model.

All results were approved for release by the U.S. Census Bureau, Data Management System Number: P-7504831 and approval number: CBDRB-FY24-ESMD002-018.

Conclusion and next steps

- This research has the potential to help improve CFS data quality and enable publishing more data, because using the recommended decision tree classifier will allow us to impute the missing mode of transportation data (7% of the cases) that cannot be fixed via rule-based imputation.
- Consider possibility to apply the same process to other missing data such as export mode of transportation.
- Consider applying the process identified in this research to other surveys with similar problems.

Acknowledgements

I would like to acknowledge and thank all of those who have helped in carry out this research.

Carla Medalia, Ph.D., Assistant Division Chief for Business Development

Berin Linfors, Branch chief of Commodity Flow Survey

Christian Moscardi, Data scientist of Business Development Staff

Chanteria Alicia Milner, Data scientist of Business Development Staff

Exploring the feasibility of imputation techniques for the Commodity Flow Survey (CFS)

QUESTIONS?

Contacts

Gritiya Tanner

Survey Statistician/Data scientist
Business Development Staff

Economic Reimbursable Surveys Division

U.S. Census Bureau

gritiya.tanner@census.gov