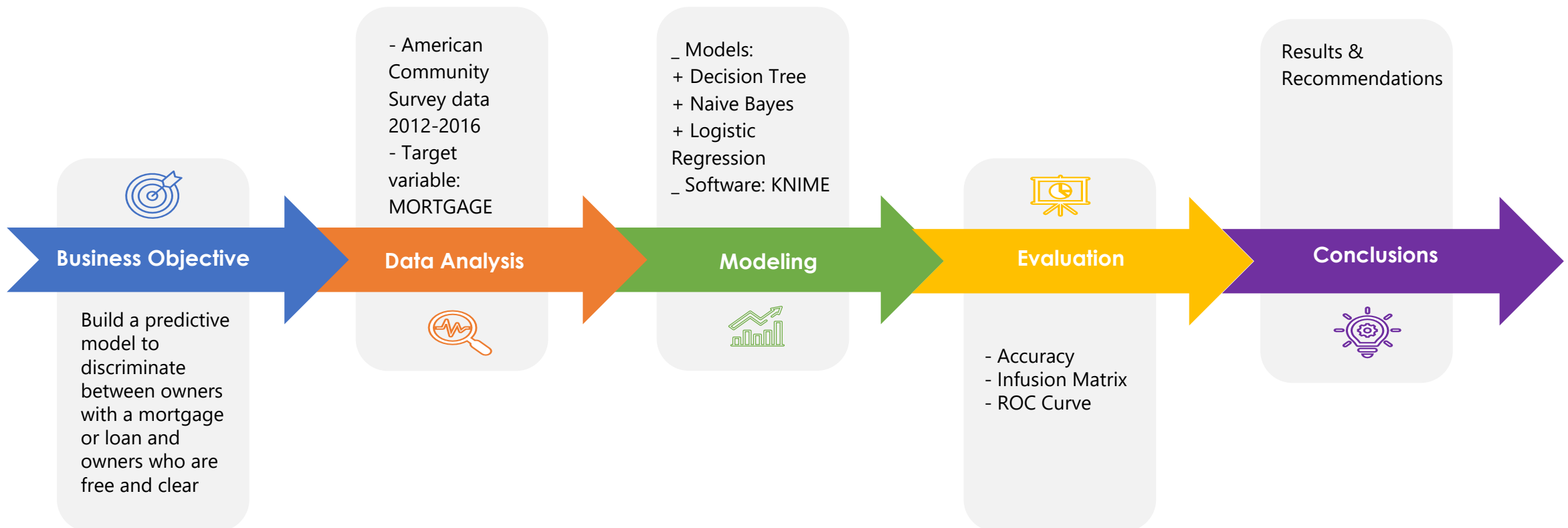# Customer Demographics Analysis
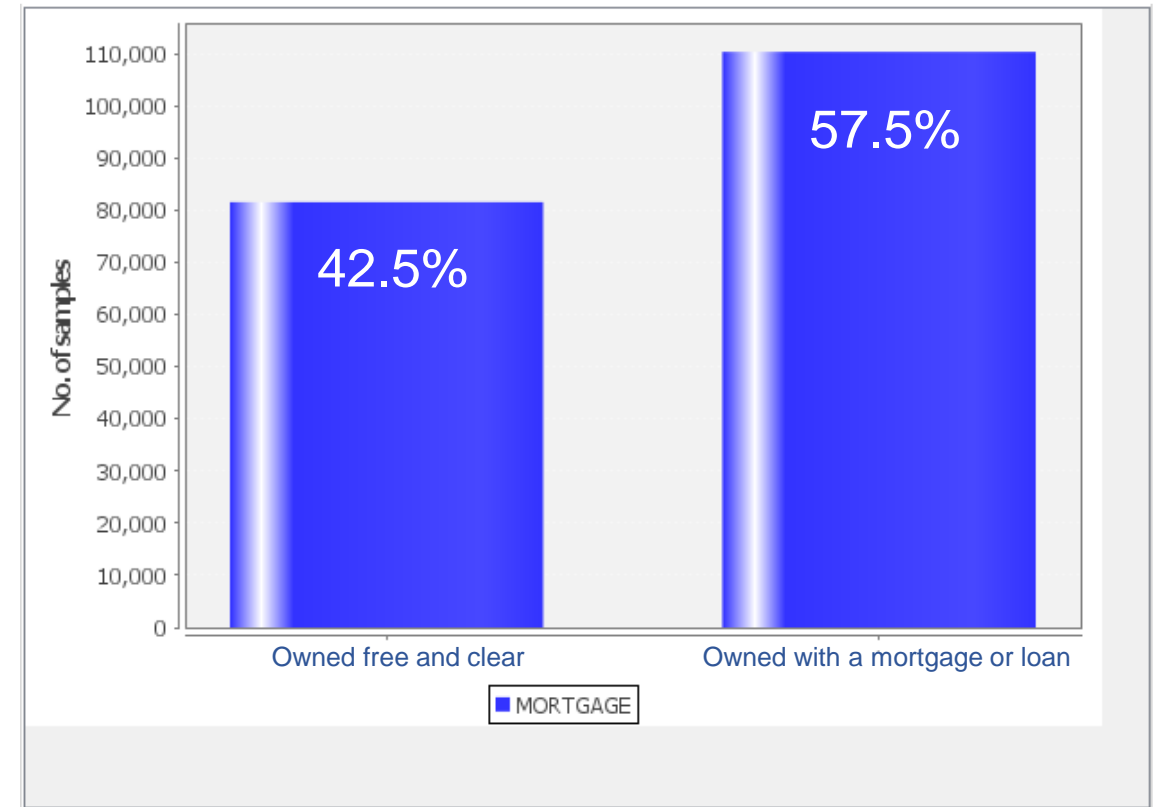## Real Estate Mortgage Company

Trinh Flynt

Problem Identification

- Identify the profiles of households who own their real estate with a mortgage or loan in the state of Pennsylvania
- Benefits:
    - Minimize acquisition costs
    - Increase marketing efficiency
    - Prepare a solid base for future marketing analysis and campaigns

Business Objective

Build a predictive model to discriminate between owners with a mortgage or loan and owners who are free and clear

- American Community Survey data 2012-2016
- Target variable: MORTGAGE

Data Analysis

_ Models:
+ Decision Tree
+ Naive Bayes
+ Logistic Regression
_ Software: KNIME

Modeling

Evaluation

- Accuracy
- Infusion Matrix
- ROC Curve

Results & Recommendations

Conclusions

Data Analysis

- American Community Survey data 2012-2016 for Pennsylvania state

- 317,416 records and 208 variables

- Derived attribute from TEN: MORTGAGE with "1" (Owned with a mortgage or loan) and "0" (Owned free and clear)

- Significant imbalance in target variable attribute

- Replace missing values

Modeling

- Modeling techniques:
    - Decision Tree
    - Naive Bayes
    - Logistic Regression

- 60% training data, 40% validation data to avoid overfitting

- Feature selection:
    - Presence of persons 60 years and over in household (R60)
    - Family type and employment status (FES)
    - When moved into this house or apartment (MV)
    - Units in structure (BLD)

- Software: KNIME

Evaluation

Confusion Matrices & Accuracy

Decision Tree

MORTGAGE...	1	0	
1	32406	11832	∧
0	8701	23741	∨

Correct classified: 56,147   Wrong classified: 20,533

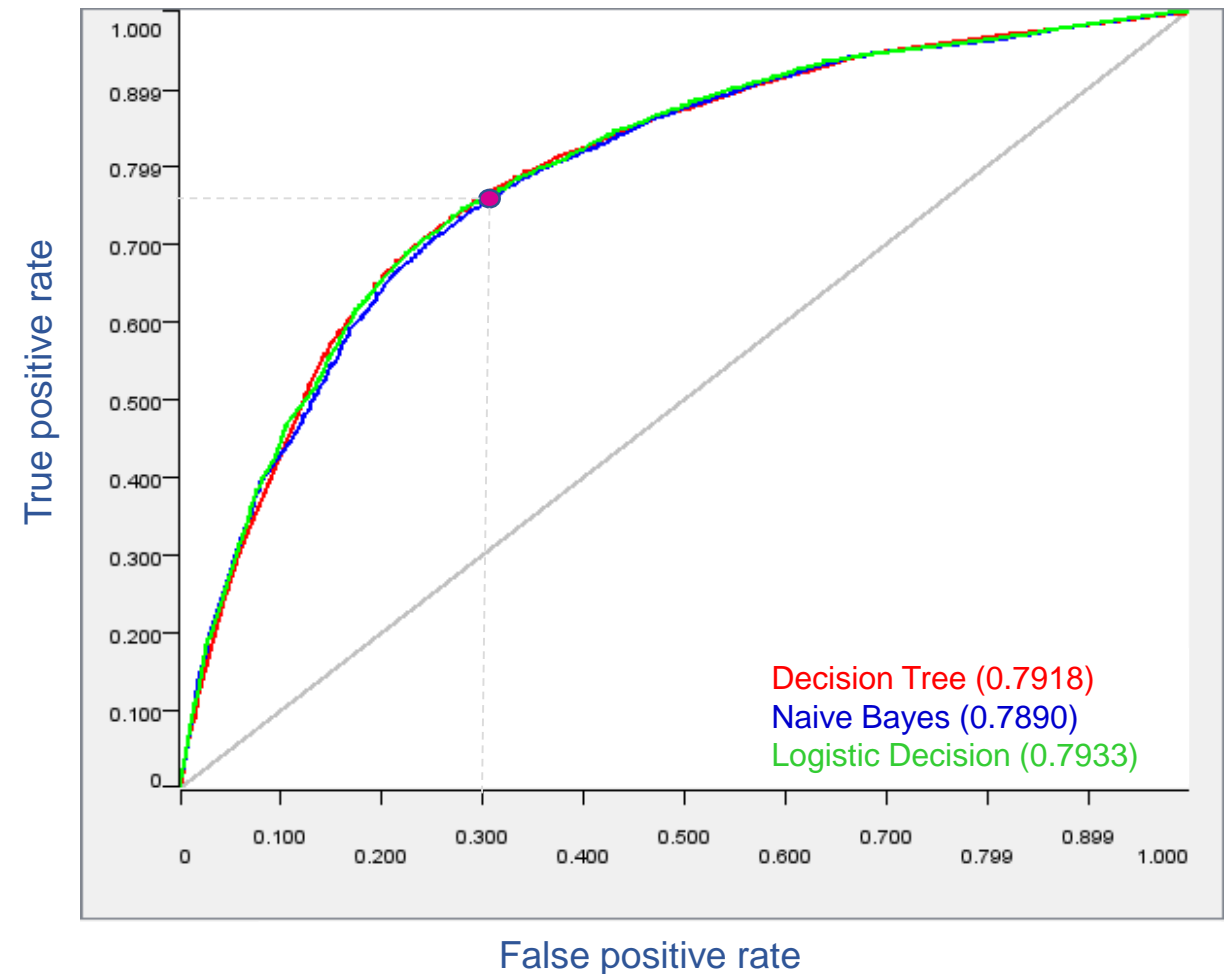Accuracy: 73.222 %   Error: 26.778 %

Cohen's kappa (κ) 0.458

Naive Bayes

MORTGAGE ...	1	0	
1	32380	11902	∧
0	9032	23450	∨

Correct classified: 55,830   Wrong classified: 20,934

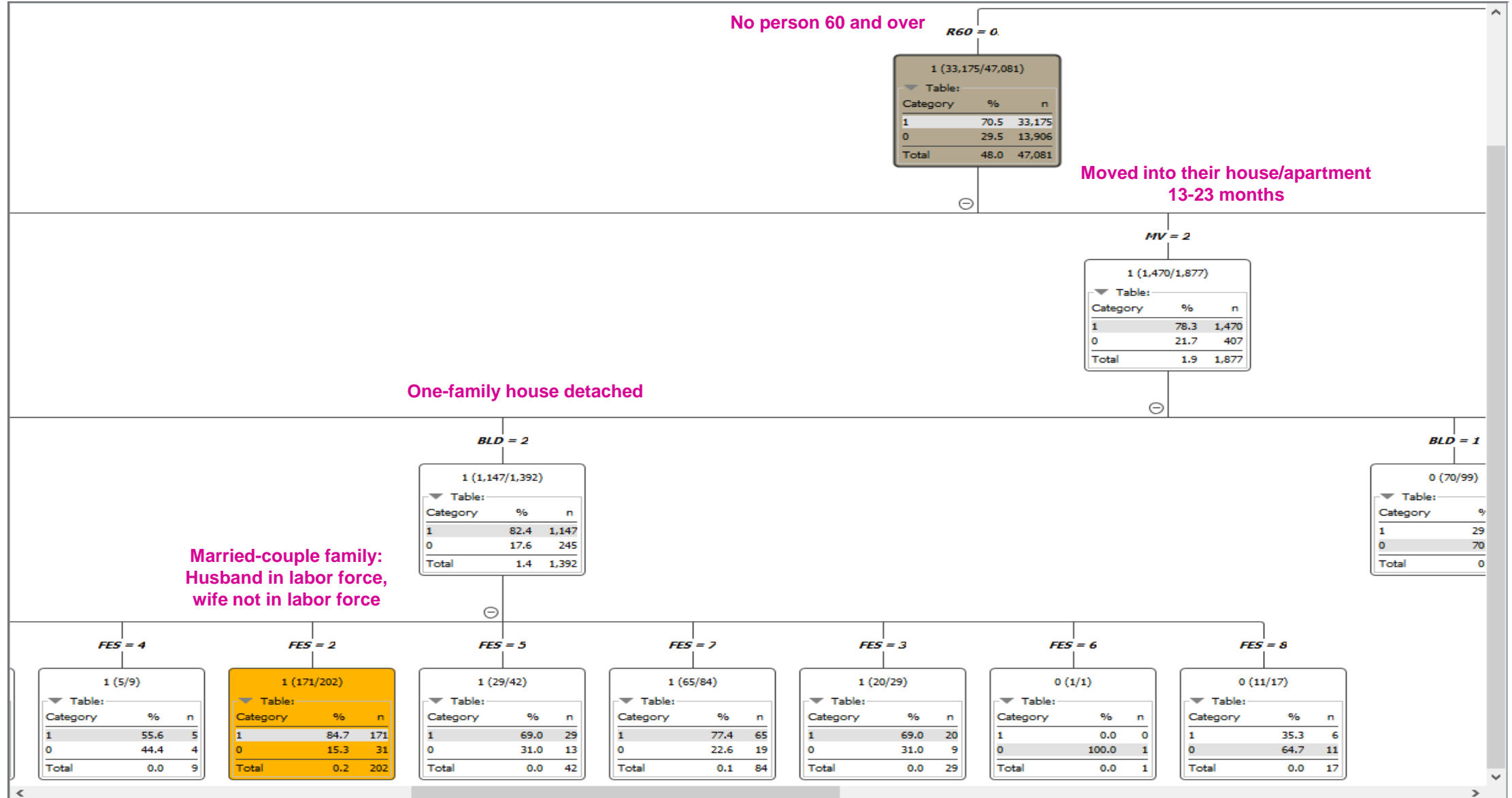Accuracy: 72.729 %   Error: 27.271 %

Cohen's kappa (κ) 0.448

Logistic Regression

MORTGAGE ...	1	0	
1	32421	11861	∧
0	8715	23767	∨

Correct classified: 56,188   Wrong classified: 20,576

Accuracy: 73.196 %   Error: 26.804 %

Cohen's kappa (κ) 0.458

ROC Curves

Decision Tree (0.7918)
Naive Bayes (0.7890)
Logistic Decision (0.7933)

True positive rate

False positive rate

Decision Tree

Conclusions

- Decision Tree and Logistic Regression model perform slightly better than Naive Bayes model

- Suggested model: Decision Tree (Highest accuracy: 73.222%, AUC = 0.7918)
    - Easy to understand, use and implement
    - Computationally cheap

- Target households who are more likely to be owners with a mortgage or loan:
    - Households without presence of 60+ years
    - Moved into their house/apartment less than 30 years
    - Not an owner of mobile home/trailer and boat, RV, van, etc.
    - At least 1 household member in labor force