Classification Model – MortgCo

By
Subatra Devi Selvaraj

Mar 20,2016

Business Problem and Approach

Business Problem and Objective:

MortgCo, the real estate Mortgage company, targets the households in Pennsylvania, who already own their real estate, but are still paying mortgage. Objective is to build a predictive model to identify profiles of households, who own their real estate with mortgage or loan, based on American Community Survey Data – 2009 till 2013.

Approach:

- **Step 1: Data Understanding and Data Preparation:**
  - Exploratory Data Analysis –Analyzed every single variable, studied the distribution, and based on the significance ranked all the variables from 1 to 5.

    1 – Most important , 5 – least important
  - Eliminated the irrelevant variables using Column Filter, treated the missing values using Missing value node.
- **Step 2: Modeling Universe:**
  - Target Variable – "TEN"(Tenure). Using Row based Filter and Rules Engine, limited universe to 2 values, 1. Owned with mortgage or loan 2. Owned free and clear. Data is partitioned and used Equal sized sampling
  - Created two models 1. Decision Tree 2. Logistic Regression

Approach

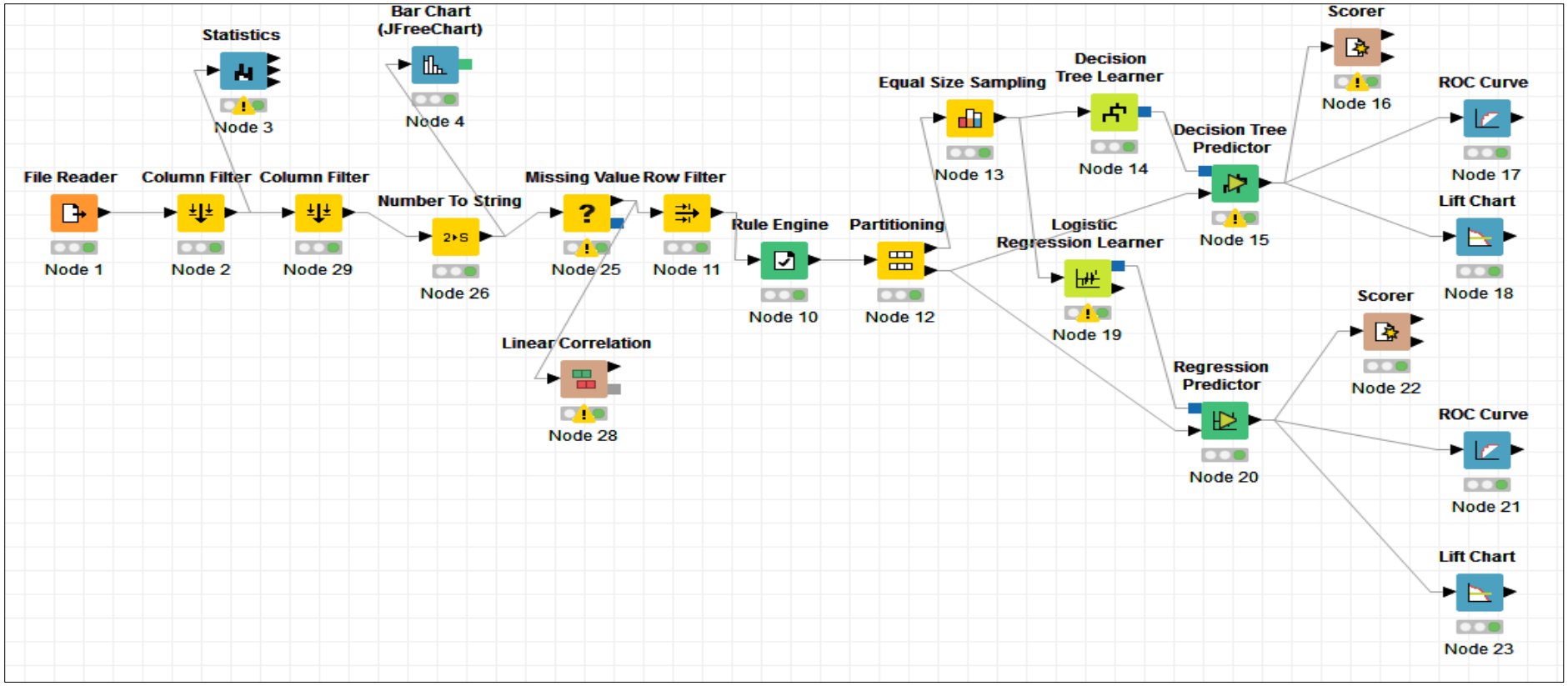
- **Step 3: Evaluation**
  - Ran the model several times, by eliminating the variables based on the **ranking**.
  - Captured the ROC, Lift curves, Accuracy stats each time for both Decision Tree and Logistic Regression.
  - Based on the Model performance, finally selected the Model with **17** variables from **205** variables.

Final List of Features by Logistic Regression:

Feature	Description
SMOCP	Selected monthly owners costs
TAXP	Property taxes
FMRGIP	First mortgage payment flag
FULP	Fuel Cost (Monthly) cost
GASP	Gas (Monthly) Cost
OCPIP	selected monthly owner costs as a percentage of household income during the past 12 months.
ELEP	Electricity (Monthly) cost
HINCP	household income (past 12 months)

Feature	Description
MHP	Mobile home costs (yearly amount)
MV	When moved into this house or apartment
VALP	property value
BLD	Units in Structure
FSMXHP	Home equity loan status allocation flag
WATP	Water (yearly cost)
HFL	House heating Fuel
YBL	Year when structure was built
R60	presence of persons 60 years and over in household

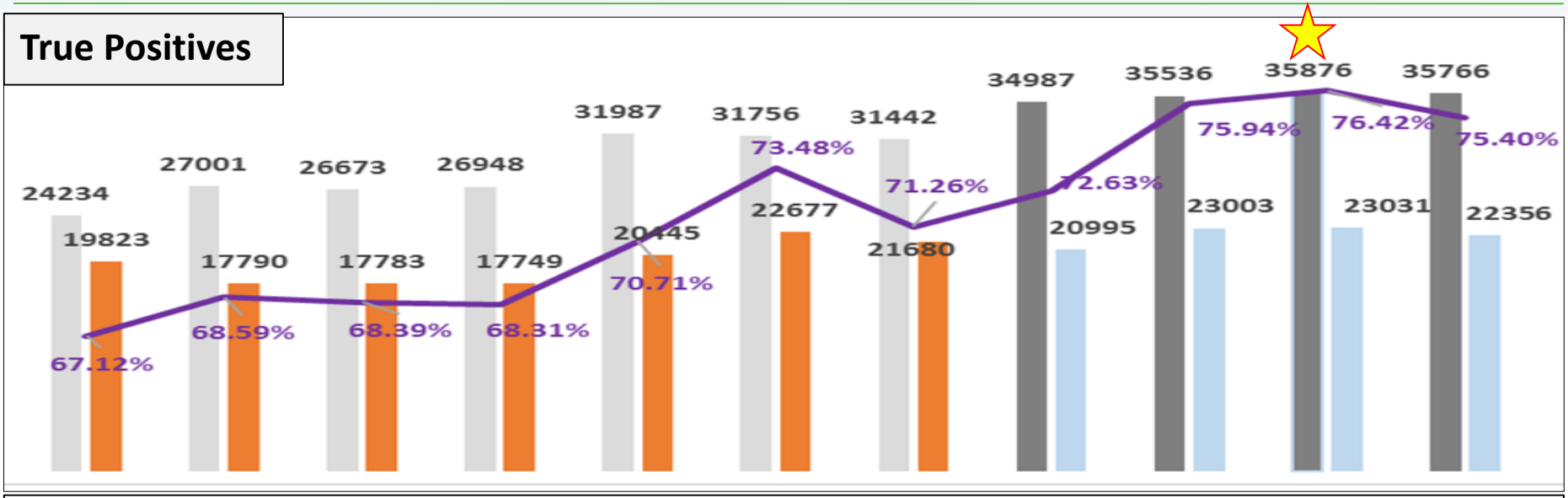
Knime Workflow



UCIrvine | Extension

Model Evaluation

True Positives

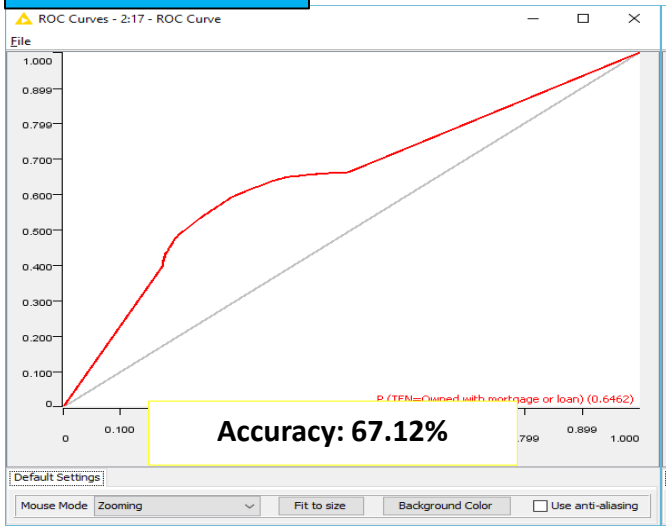


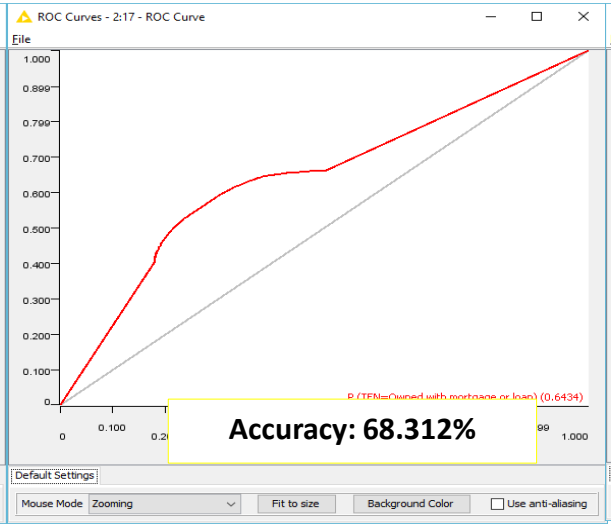
Selected Final Logistic Regression Model Stats	Predicted Owned with mortgage or loan	Predicted Owned free and clear	Sensitivity	Specificity	F-measures	Accuracy
Owned with mortgage or loan	35876	8693	0.790935	0.72598	0.79788275	0.76420222
Owned free and clear	9483	23031	0.72598	0.790935	0.71705221	

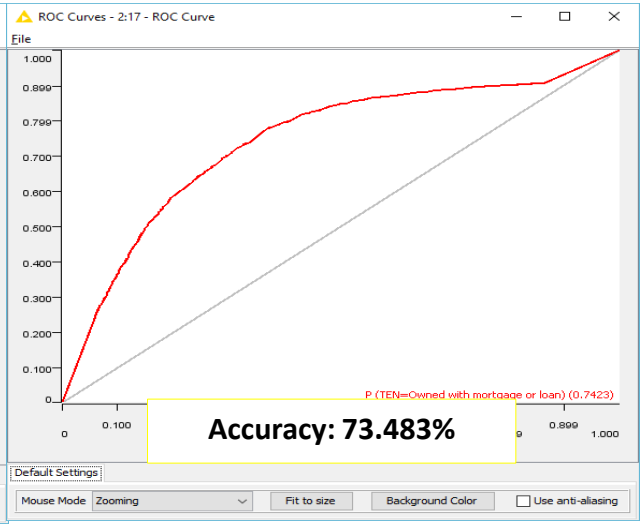
UCIrvine | Extension

ROC Curves

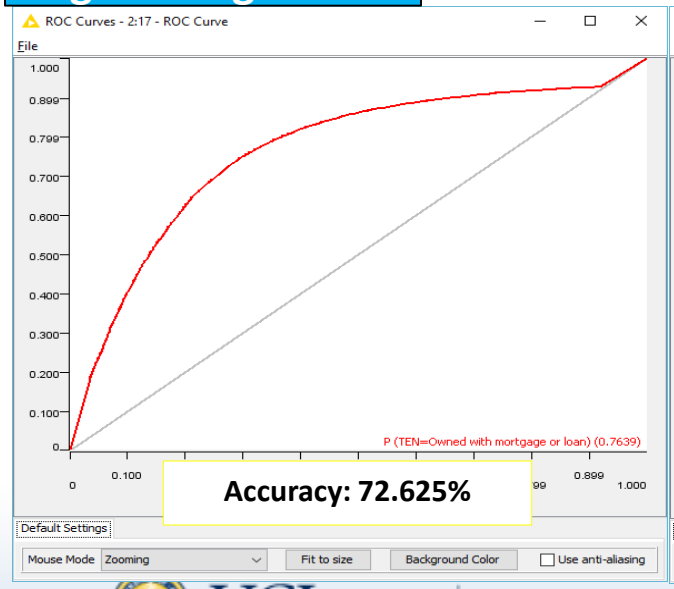
Decision Tree

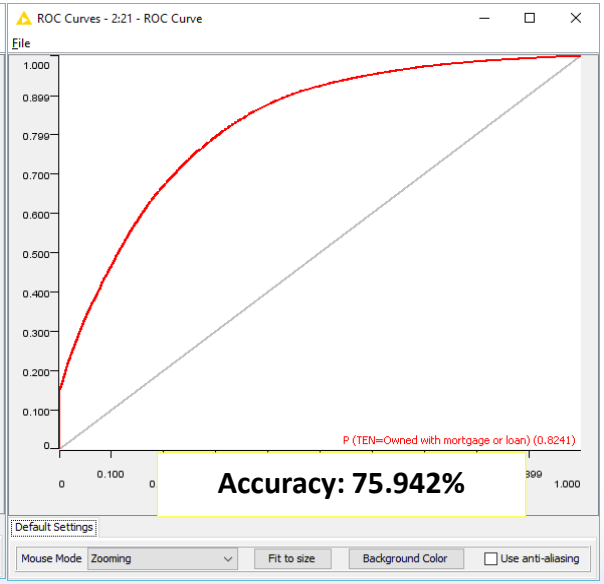


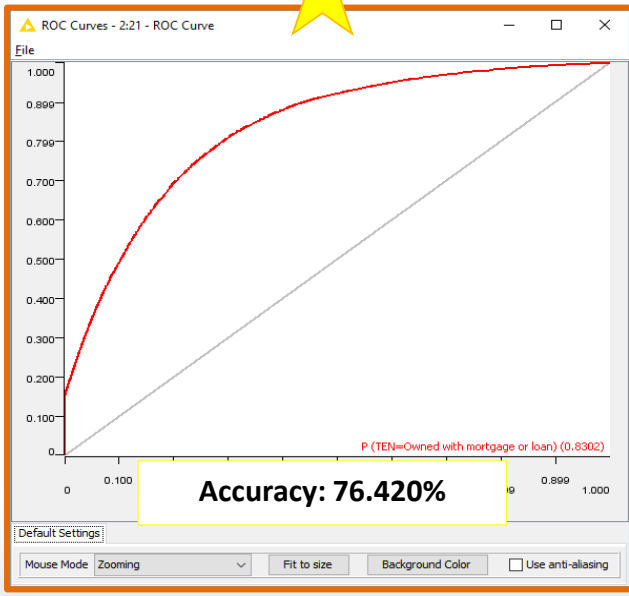




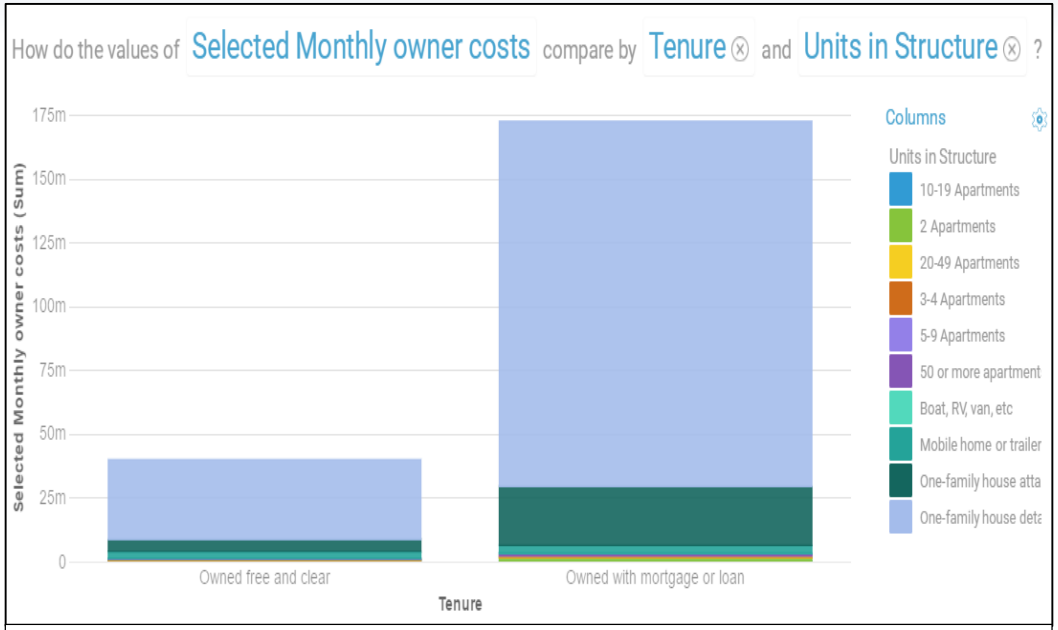
Logistic Regression

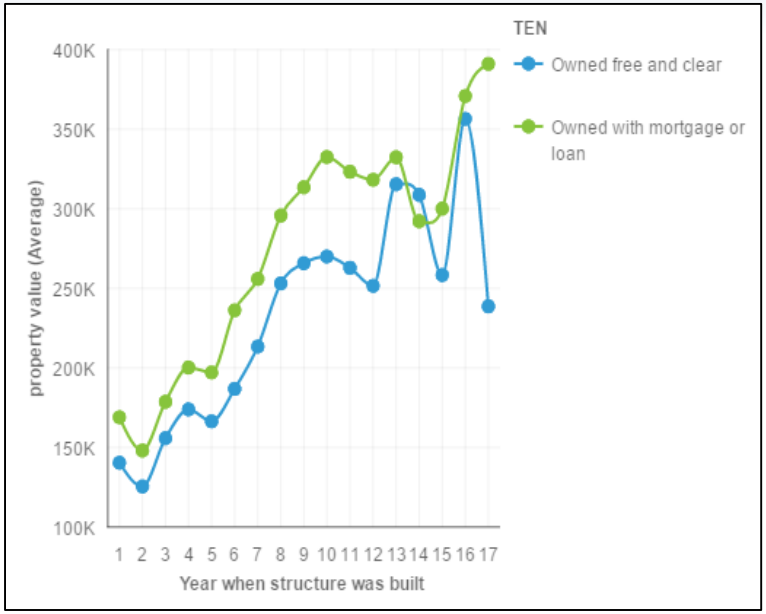


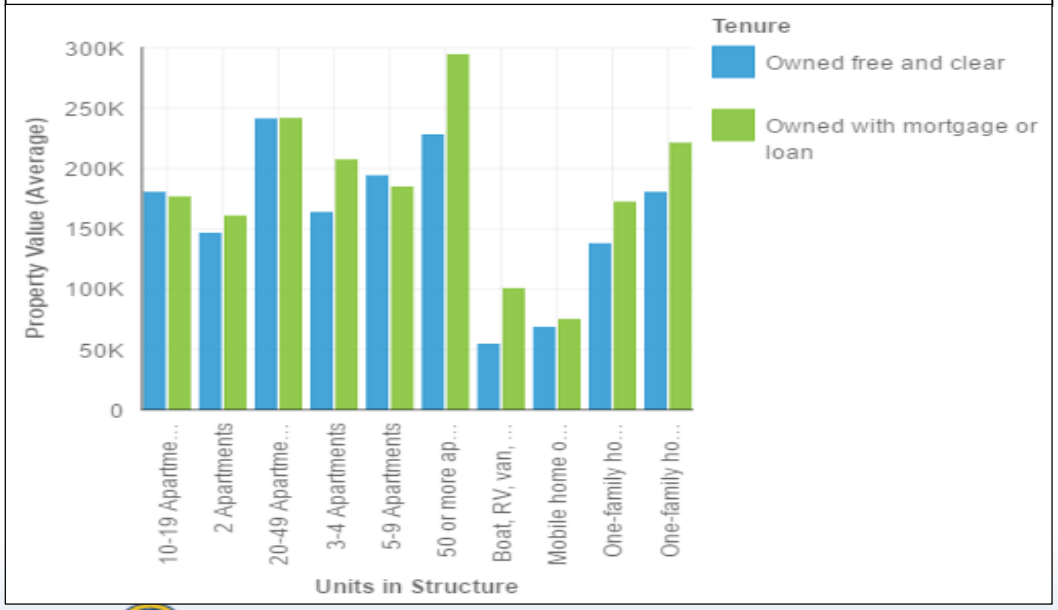


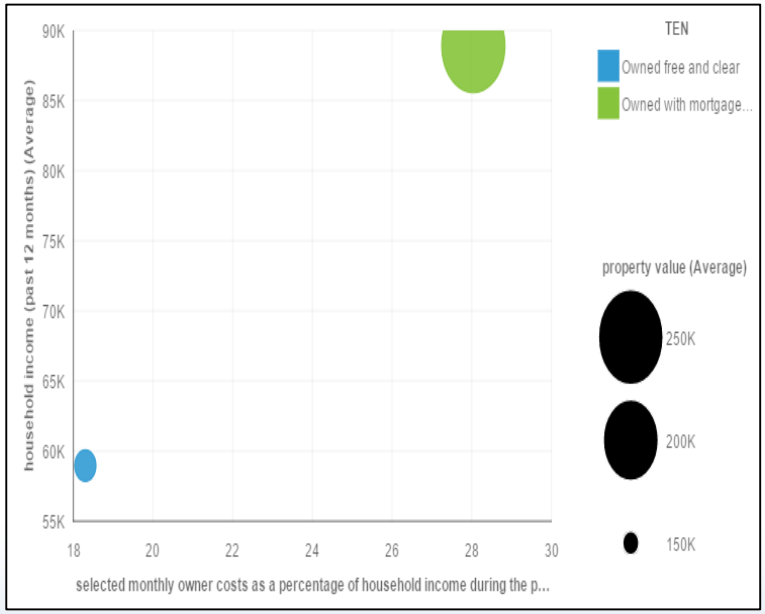


Significant Features Vs Target(TEN)









Executive Summary

Conclusion

- Logistic Regression Model is best suited for this classification. The 17 features identified (Slide 3) gives the best results in prediction.

Mortgco can target the population below

- R60 (Presence of persons 60 years or over in household) = 0 (No persons > 60 years)
- MV (When moved into this house) < 29 years
- Units in Structure – 2 to 8 (One-family house, Apartments)

Significant Variables:

BLD	Units in Structure
ELEP	Electricity (Monthly) cost
GASP	Gas (Monthly) Cost
HFL	House heating Fuel
YBL	Year when structure was built
MV	When moved into this house or apartment
R60	presence of persons 60 years and over in household (unweighted)
FMRGIP	First mortgage payment includes fire, hazard, flood insurance allocation flag
FSMXHP	Home equity loan status allocation flag