FIG WORKING WEEK 2023
28 May - 1 June 2023 Orlando Florida USA
Protecting Our World, Conquering New Frontiers

Using Machine Learning to Create High Performing Models for Automated Valuation

May 29th, 2023

Organized By NSPS FIG INTERNATIONAL FEDERATION OF SURVEYORS
Diamond Sponsors esri THE SCIENCE OF WHERE Trimble

Meet the Speakers

Luke Jorgensen

Lee County Property Appraisers

Business Data Analyst

Joshua Jorgensen

Lee County Property Appraisers

Data Scientist

Introduction

- Mass Appraisal is key to being able to quickly and fairly appraise large volumes of properties
  - One of the difficulties is ensuring equity and fairness of the Mass Appraisal values
- Linear/Logistic Regression has been the gold standard for creating Mass Appraisal models
  - Ability to learn from very large data sets and provide accurate and unbiased predictions
  - Model interpretability
- Machine Learning models have made great strides in the past decade
  - Machine learning models offer high levels of accuracy
  - Explainable AI allows for the interpretation

Linear Regression Models

- Use statistics to determine the contribution and effect of predictors to predict the value of a parcel
- Using coefficients (β), models illustrate the effect of a predictor on an outcome
  - Ex. For one unit increase in base area, the parcel's value increased by $10.02
- Regression models do have restrictions

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Machine Learning Models

- There has been significant advancements in the field of machine learning over the past decade
  - Consist of algorithms that try to learn latent patterns and relationships from data without hard coding fixed rules
  - Allows the models to capture non linearities in the data
  - Many ML algorithms can handle non independent features and even imbalanced data

Data For the Model

- All models were trained and tested on Market Area 02
  - Vast swathes of housing in Estero, FL and Bonita Springs, FL
- 10 years of sales
  - Totaling 24,556 sales

Lee County
Southwest Florida

Model Construction

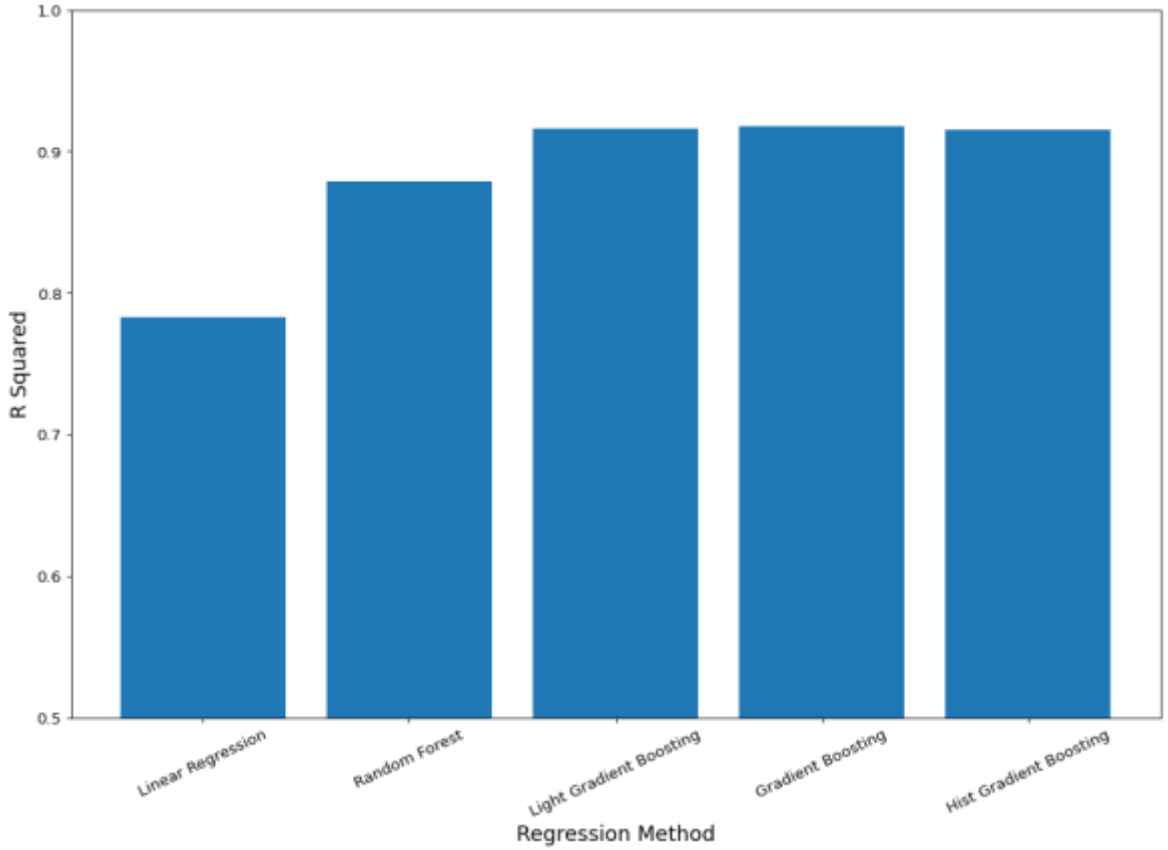
A model is considered high performing if it has a high level of accuracy and utilizes as few independent variables/predictors as possible

- Less than 20 features were used to predict the parcel values
  - Same predictors were used to train and test each model
  - Features included base area, age, improvement code, quality, land square footage, etc.

- Models were compared based on the performance metrics of R-squared, price related differential, and coefficient of dispersion
  - Allowed for gauging performance as well as fairness and equity of the models

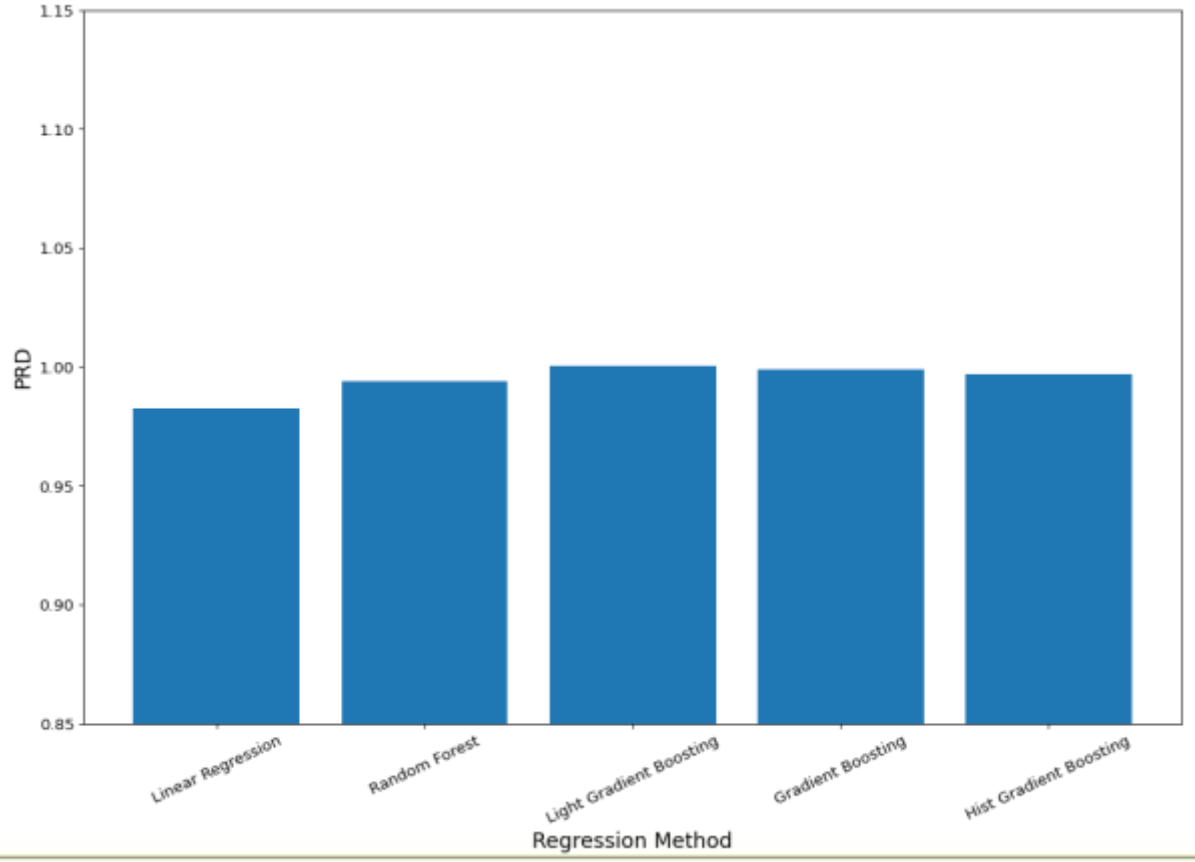
R Squared

- ✓ Linear Regression: 0.783
- ✓ Random Forest: 0.879
- ✓ Light Gradient Boosting: 0.916
- ✓ Gradient Boosting: 0.918
- ✓ Histogram Gradient Boosting: 0.916



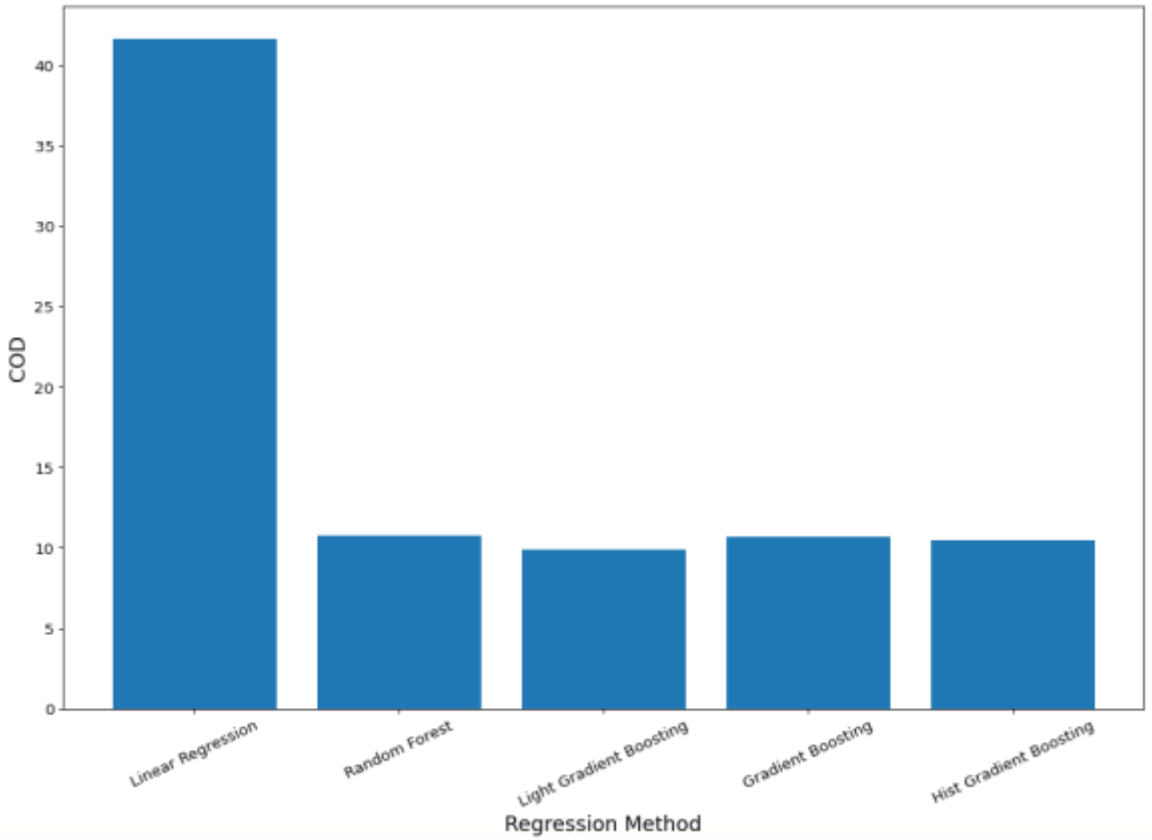
Price Related Differential

- ✓ Linear Regression: 0.983
- ✓ Random Forest: 0.994
- ✓ Light Gradient Boosting: 1.00
- ✓ Gradient Boosting: 0.999
- ✓ Histogram Gradient Boosting: 0.997



Coefficient of Dispersion

- ✓ Linear Regression: 41.62
- ✓ Random Forest: 10.75
- ✓ Light Gradient Boosting: 9.90
- ✓ Gradient Boosting: 10.65
- ✓ Histogram Gradient Boosting: 10.49



Hyperparameter Tuning

- Parameters are what the "model" uses to make predictions

- Hyperparameters are what a machine learning model learns
  - Come with default hyperparameters, they might not be optimal for your given process
  - Models can consist of a multitude of hyperparameters
  - Some of the hyperparameters can take on an infinite number of values

- Optuna is an automatic hyperparameter optimization software framework
  - Uses Bayesian Statistics to crawl through a custom search space of hyperparameters to optimize performance
  - Control metrics to increase performance

Improvements on Machine Learning Models from Optuna

LightGBM

Metric	Value
R Squared	0.916
PRD	1.00
COD	9.90

Metric	Value
R Squared	0.924
PRD	0.999
COD	9.19

Gradient Boosting

Metric	Value
R Squared	0.918
PRD	0.999
COD	10.65

Metric	Value
R Squared	0.919
PRD	1.00
COD	9.88

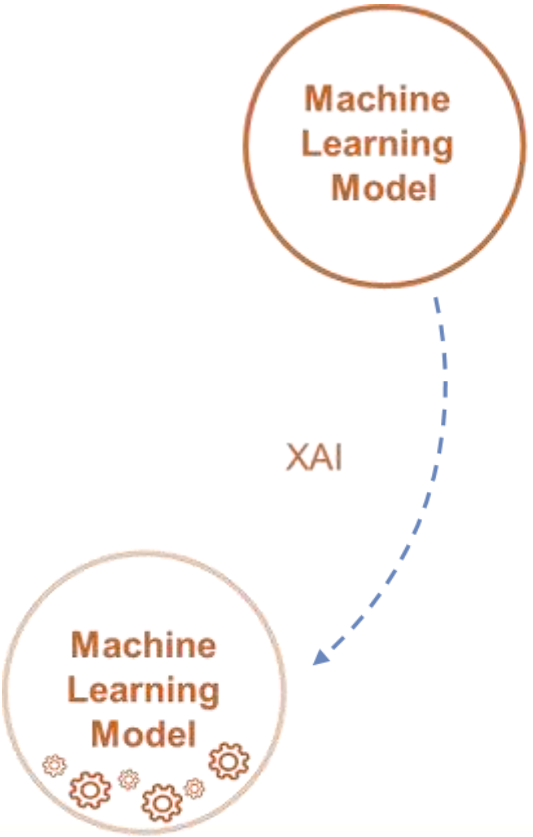
What stops some people from adopting Machine Learning?

- The strength of using algorithms without hard coding fixed rules also creates a weakness
  - Explaining how these models work always poses its own set of challenges
  - They change dynamically depending on what data point you are predicting; making them harder to explain
- Many individuals think of Machine Learning models as a black box
  - In turn, people do not trust the predictions provided by the model

How do you overcome this obstacle?

- Explainable AI (XAI) helps us to understand how a model is making its predictions
  - Ex. What features are positively or negatively impacting the outcome of a prediction?

- With Explainable AI, we can answer the following questions:
  - Why does the model predict that result?
  - What are the reasons for this prediction?
  - What are the strongest contributors to the prediction?
  - How does the model work?

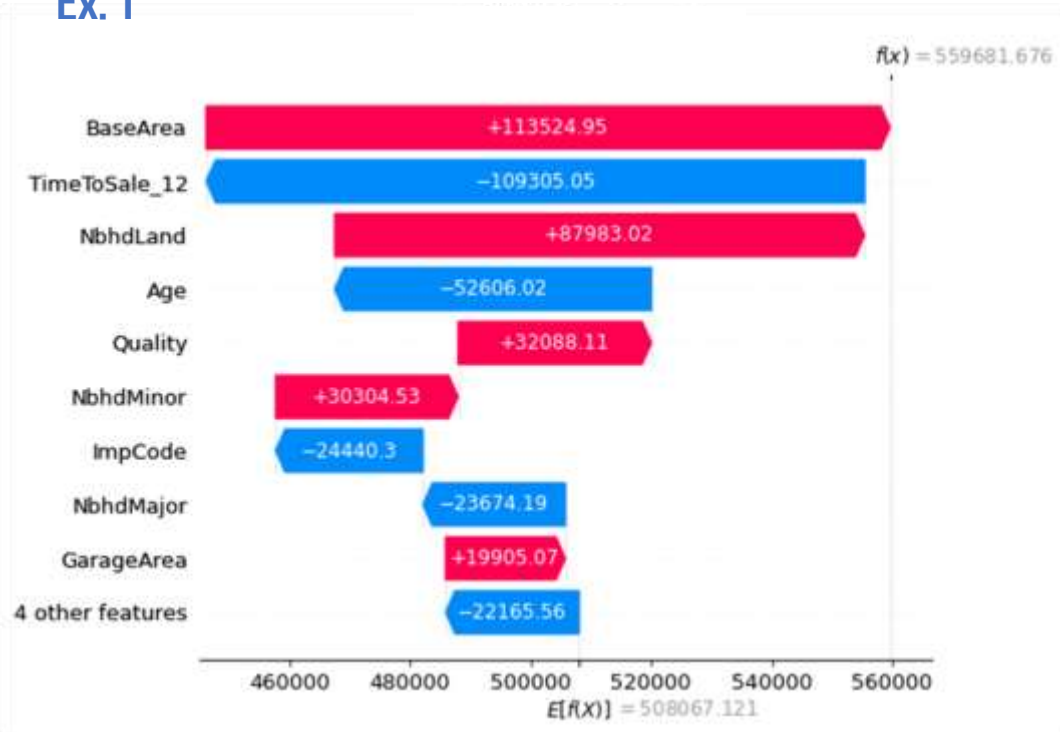


SHAP

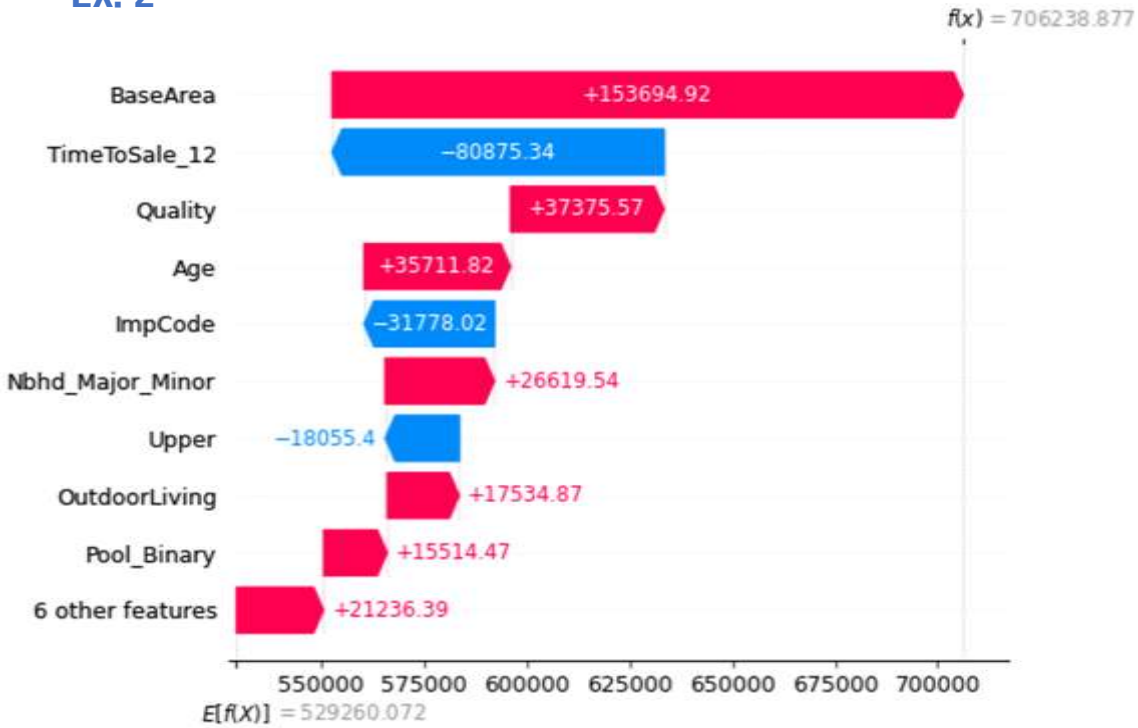
- SHAP stands for **SH**apley **A**dditive ex**P**lanations
- SHAP is a method for explaining Machine Learning by using concepts of game theory to reverse engineer the output of any predictive model
- SHAP is quantifying the contribution that each feature brings to the prediction made by the model
  - By using the outcome of each possible combination of features it determines the importance of a single feature
- SHAP offers unified global and local model interpretability

Local Explanations

Ex. 1

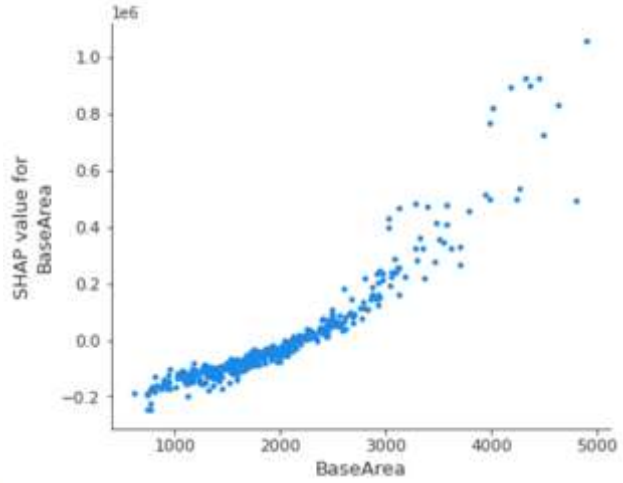


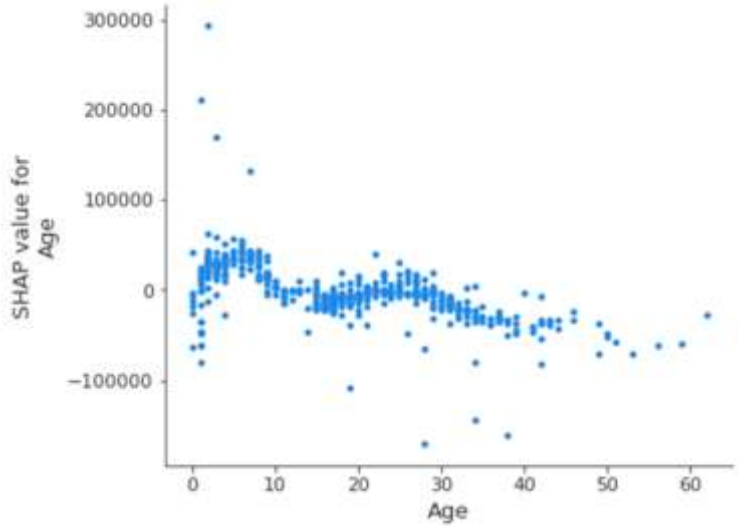
Ex. 2



Global Explanation

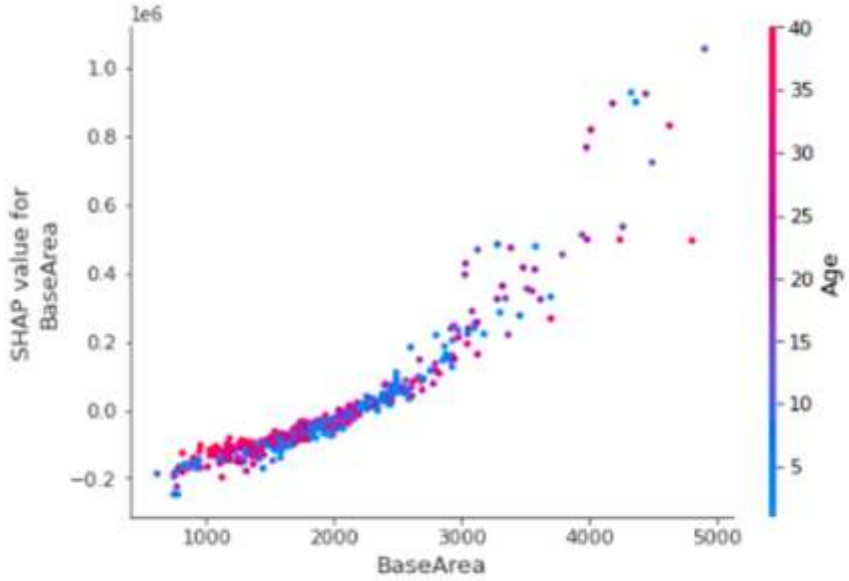
- SHAP can also visually represent the impact any predictor is making on the model for the entire range of values for that predictor
  - Allows for discussion with appraisers
  - Extraction of rate curves





Interaction of Features at a Global Level

- Many times, features can impact one another
  - Interaction plots let us see the importance of one feature to the model
  - Visually shows its relationship to another feature



Takeaways and Questions

- Employing machine learning models addresses many of the limitations of linear regression models
- Machine learning models have not only outperformed the linear regression models, but they also maintained appraisal vertical and horizontal equity where the linear model failed to do so
- Explainable AI techniques such as SHAP were applied to the model in order to provide a high level of "explainability" to the outcomes that capture nonlinearities and other nuances

Thank You!

Luke Jorgensen
- Email: jorgensenl@leepa.org

Joshua Jorgensen
- Email: jorgensenj@leepa.org