

Predicting Diamond Prices

Mychael Gray

**Data Science Intensive Capstone Project, November 1st 2021
Cohort**



Thanks to Springboard mentors



A Wang Ang, Principal Data Scientist



Michelle Jorgensen, Springboard Student Advisor

The Problem

According to Bloomberg, “While Americans are buying more diamond jewelry than ever before... The lower prices and a glut ... have pushed the global diamond trade into crisis”.

Americans spent:

36+ billions

Demand growth rate:

4.5% (in America which accounts for 50% of sales)

A Lost Decade

Wholesale diamond prices are floundering

✍ PolishedPrices.com Diamond Index



Source: PolishedPrices.com

What opportunities exist for interested parties of diamonds to predict the value of a particular diamond better than the historical prices?

Who might care?

Investors



Businesses



Buyers/Consumers



...and many more...

What factors might affect prices?

- Prices: set by sellers or wholesalers
- Carat or weight
- Cut or quality
- Clarity
- Color

kaggle



Data Source

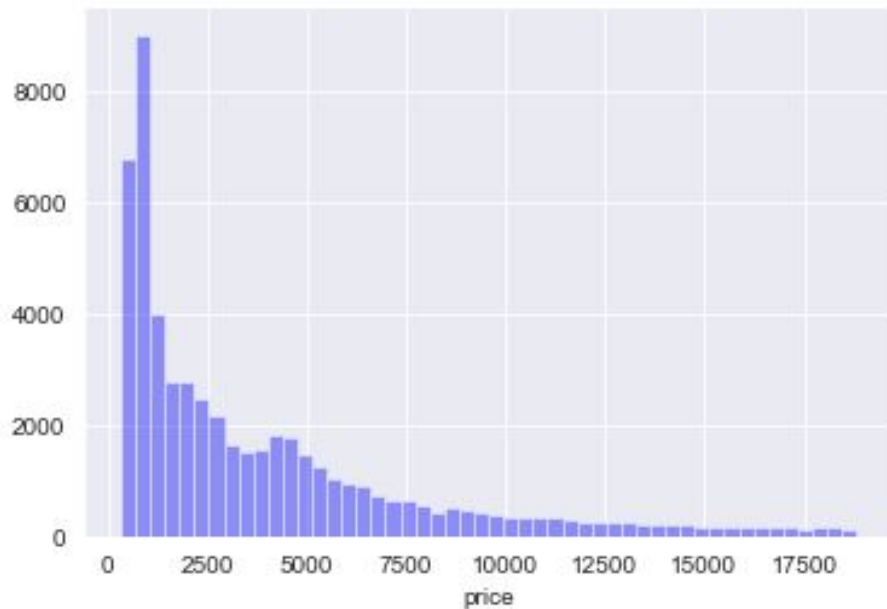
- Technical Factors:

- Table
 - width
 - length
 - depth

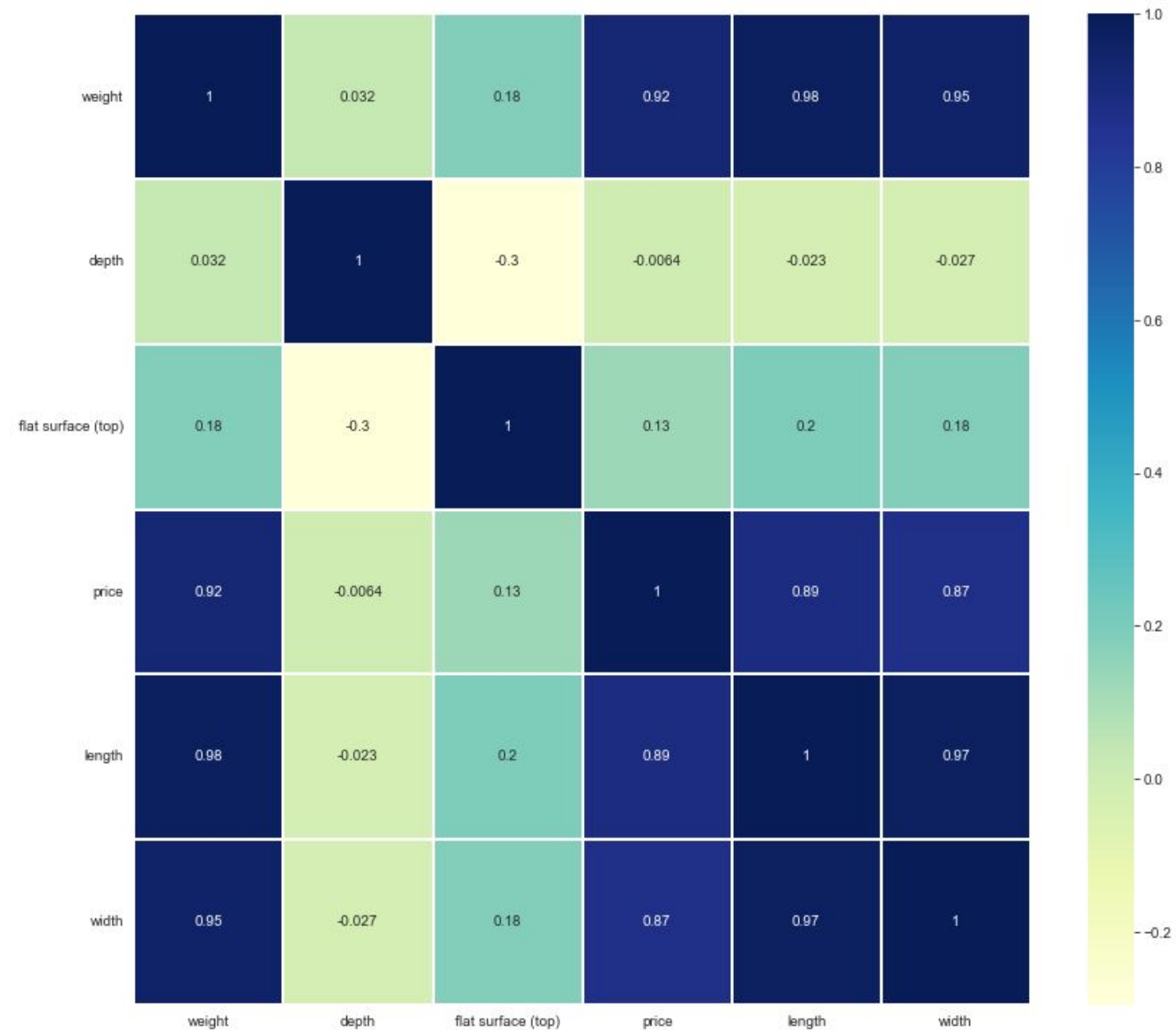
Industry



Some Historical Indicators



Initially when exploring the data, we could see that the lowest prices start at 326.00 in US dollars with 2 observations of that amount and can raise all the way up to 18823.00 in US dollars with 1 observation at that amount.



Modeling Overview

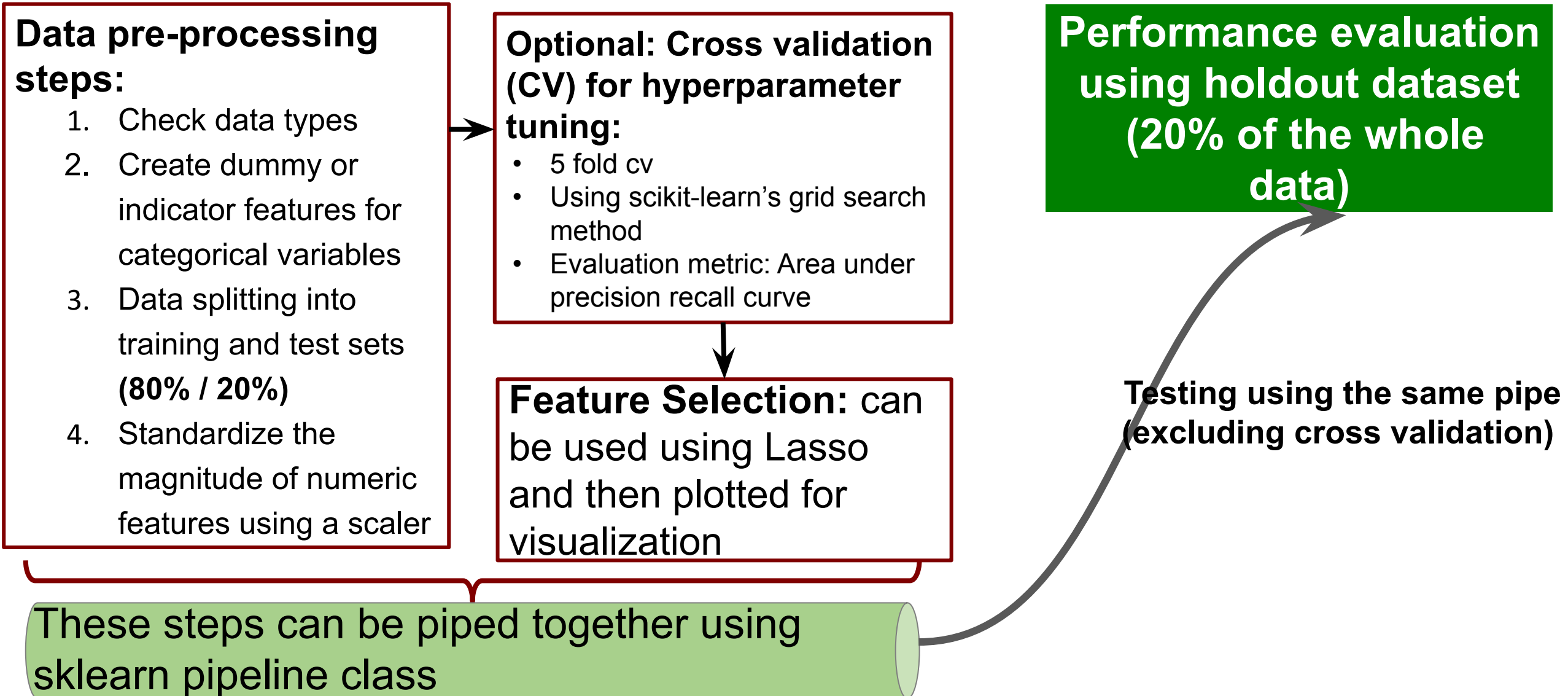
Type: Supervised learning

Regression Model: Predicting Price

No missing data & feature names updated

Tools: Python 3, scikit learn, and matplotlib

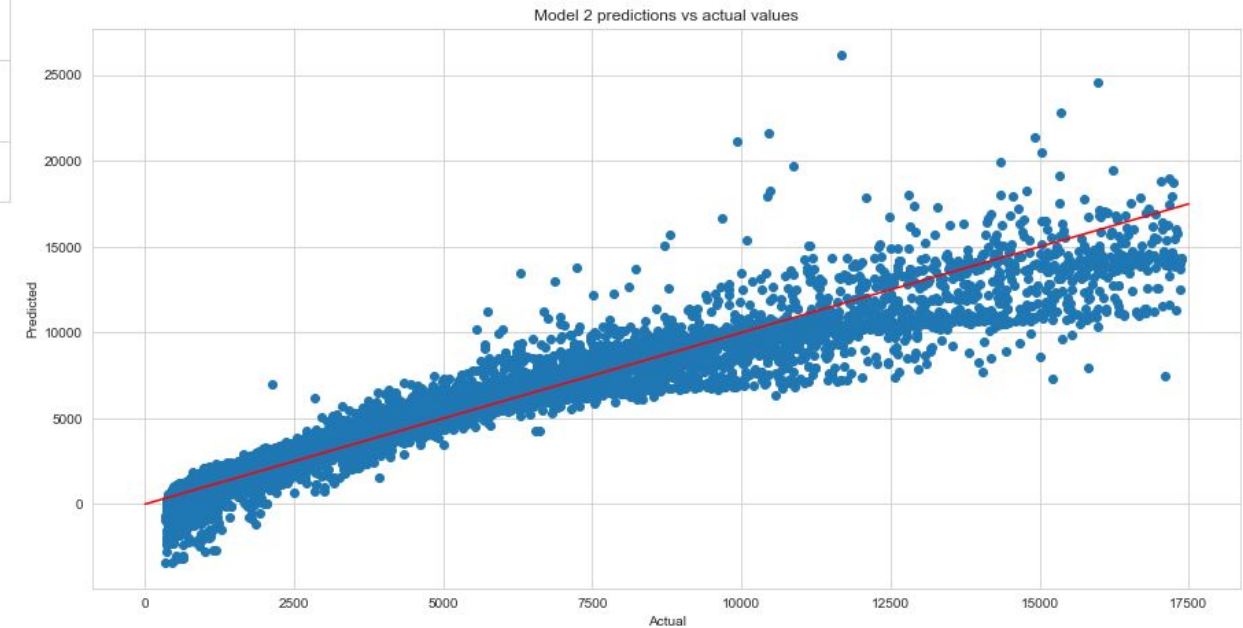
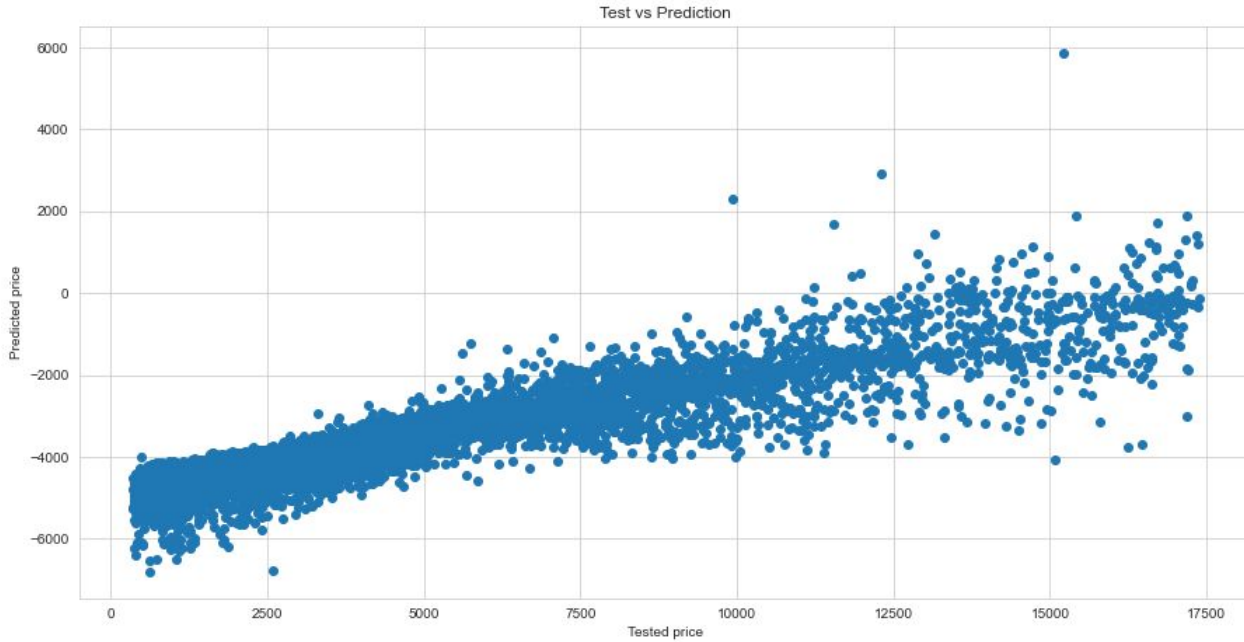
Modeling Steps



Regression Algorithms Used (Four Models):

1. Linear Regression – using sklearn
2. Linear Regression (Ordinary Least Squares) – using statsmodels
3. Ridge Regression – using sklearn
4. Lasso Regression – using sklearn

Model Comparisons



Logistic Regression using **sklearn** and **Logistic Regression** using
Ordinary Least Squares

Details on the Best Model (ET)

Making a Ridge Regression model: Third model

Sklearn has a `Ridge()` function built into the `linear_model` module.

```
4]: from sklearn.linear_model import Ridge

# Split data if not split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state=42)

# Initialize ridge model, normalize=True parameter assures all variables are scaled.
ridge = Ridge(alpha=0.1, normalize=True)

# We can use the scaled data from above to fit
ridge.fit(X_train, y_train)

# Make predictions
ridge_pred = ridge.predict(X_test)

# Score predictions
ridge.score(X_test, y_test)

4]: 0.8817195230671278
```

- By using the Lasso Regression Model we were able to increase the predictions of the predictive power to about 91%.

Making a Lasso Regression model: Fourth model

Sklearn has a `Lasso()` function built into the `linear_model` module.

```
from sklearn.linear_model import Lasso

# Split data if not split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state=42)

# Initialize ridge model, normalize=True parameter assures all variables are scaled.
lasso = Lasso(alpha=0.1, normalize=True)

# We can use the scaled data from above to fit
lasso.fit(X_train, y_train)

# Make predictions
lasso_pred = lasso.predict(X_test)

# Score predictions
lasso.score(X_test, y_test)

0.909819917007543
```

While the first model scored the highest the fourth model is the best to used for added complexity.

- By using the Ridge Regression Model we were able to increase the predictions of the predictive power to about 88%.

Assumptions, Limitations and Disclaimers

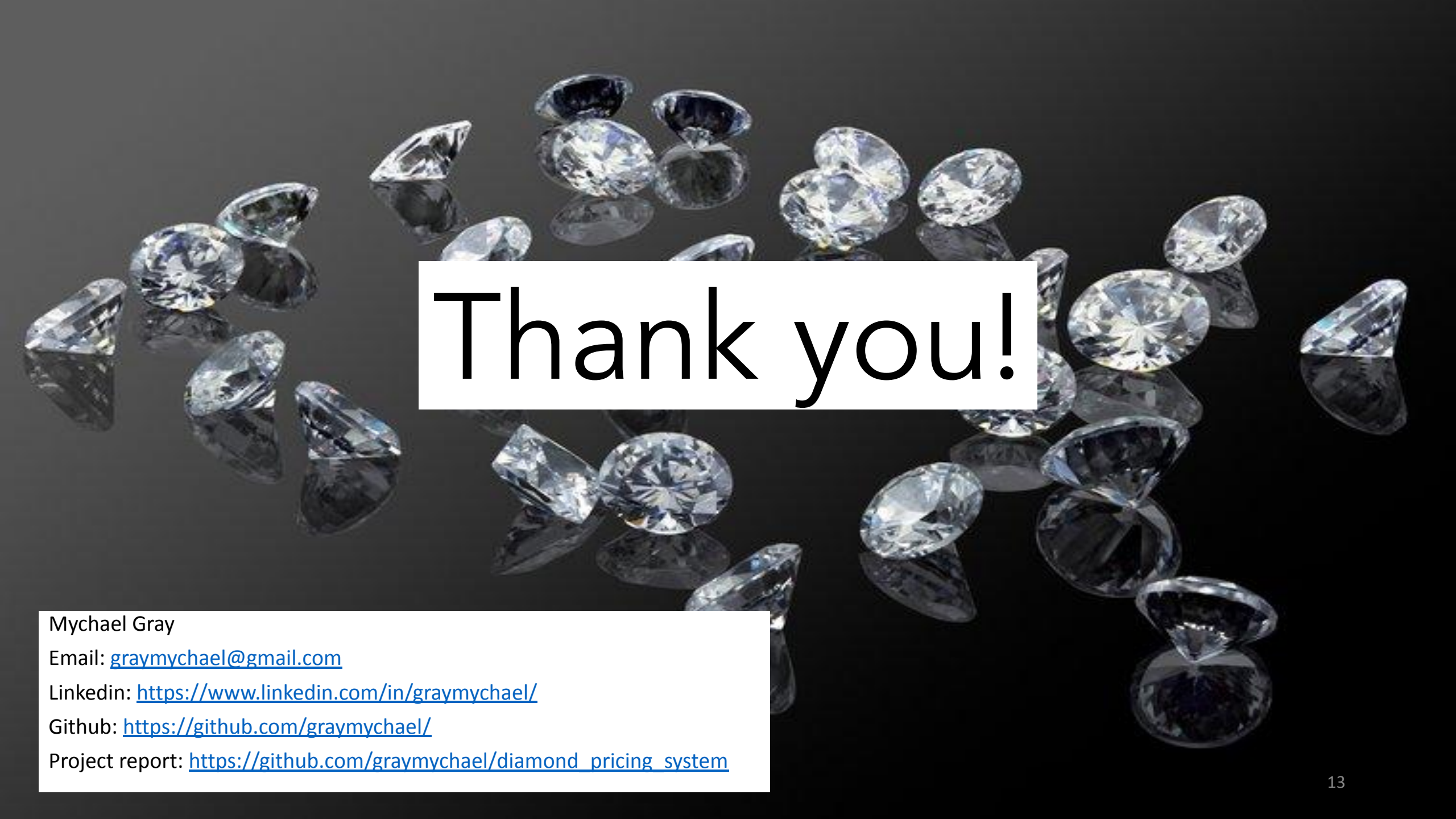
- ❑ We assume that all observations are independent, are assumed to be unrelated diamonds
- ❑ Used only 53940 observations of diamonds and past data that is not current
- ❑ The model will behave poorly if we try to predict prices of diamonds that are scheduled too far in future
- ❑ The diamonds data does not contain all data (several sources can be included in the future)

More Ideas to Improve the Model in Future

- Engineer more features related with diamonds such as customer feedback, customer surveys, environmental/social impact (blood diamonds), etc.
- Answer which diamonds are more volatile in price or more stable and why?
- Use social network and news media to extract sentiments about diamonds and stock market/economy health to get more features

Conclusions

- All sources of datasets contributed to the predictive power of the model.
- Out of 5 supervised classification models, the Extremely Randomized Trees provided the best results.
- Out of 67 features, we used only 31 features for the best model with 12 from the flight data, 16 from the weather data and 3 from the flight historical performances data (which we engineered).
- With 50%-50% splitting, the test data set gave ROC AUC = 0.89.



Thank you!

Mychael Gray

Email: graymychael@gmail.com

Linkedin: <https://www.linkedin.com/in/graymychael/>

Github: <https://github.com/graymychael/>

Project report: https://github.com/graymychael/diamond_pricing_system