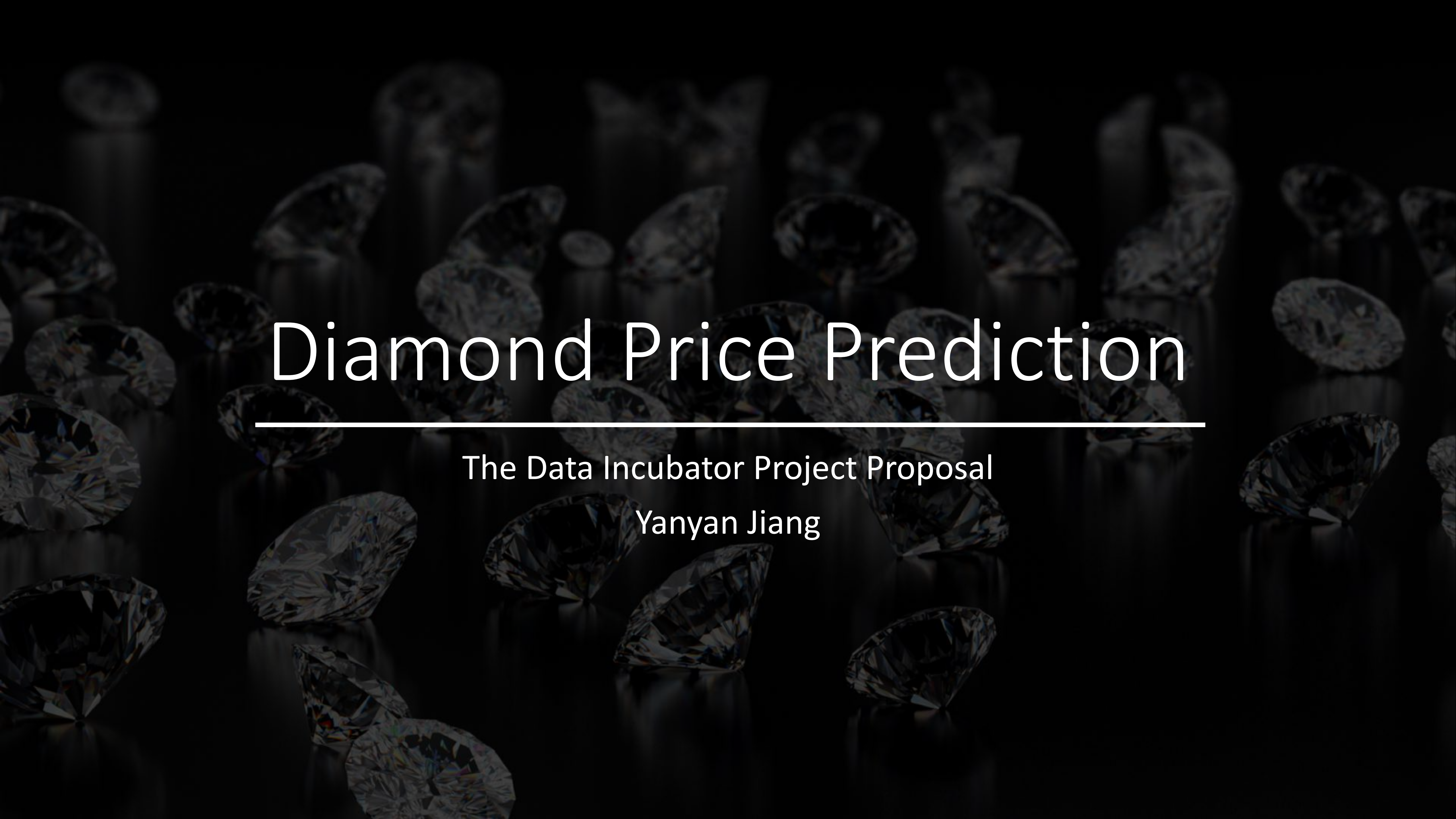


Diamond Price Prediction

The Data Incubator Project Proposal

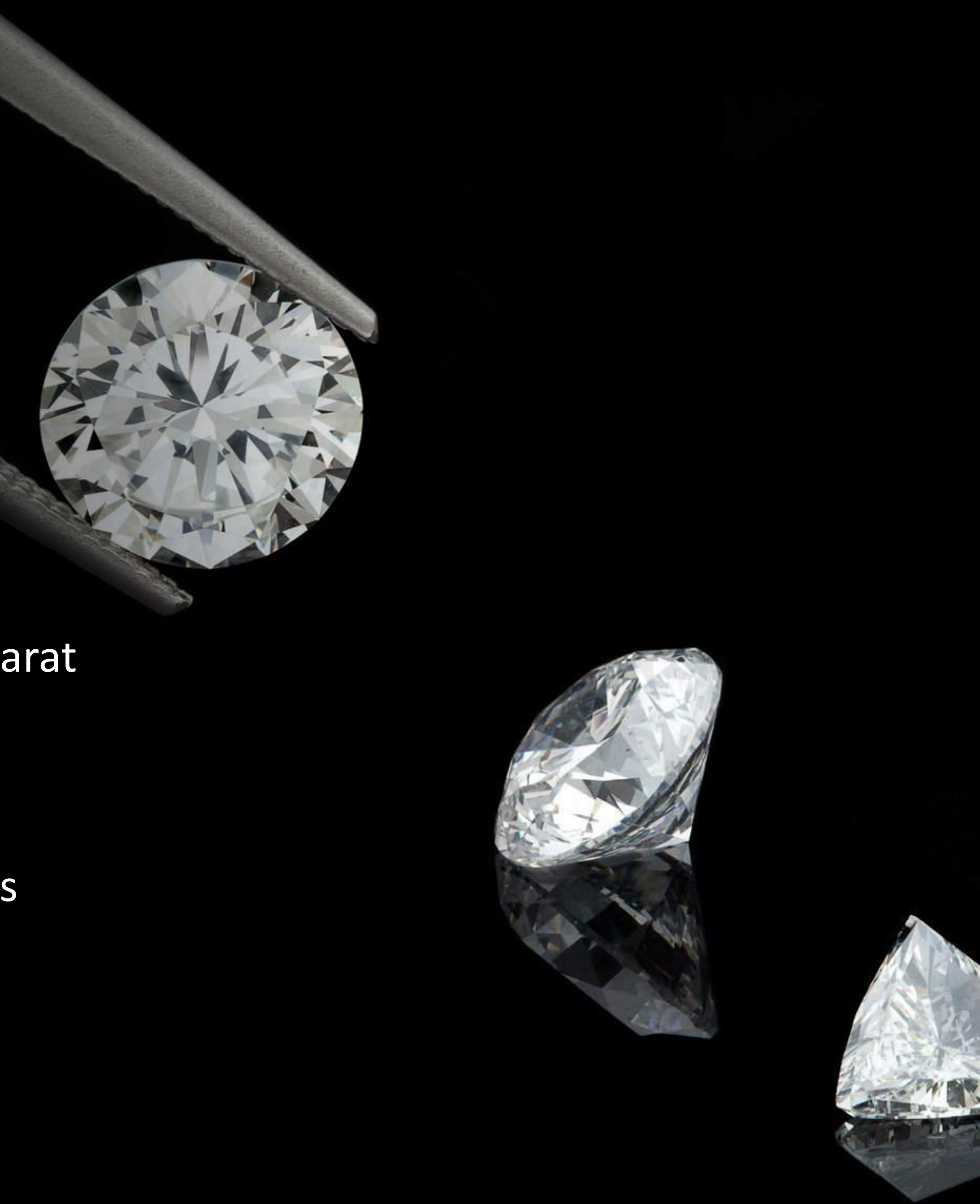
Yanyan Jiang

Introduction

- Most people (80%, 2019 survey) buy a diamond engagement ring
- Build a model that can predict diamonds' market prices and give customers advices on how to pick the best diamond without overpaying too much

Dataset Preparation

- Collect data from the official BlueNile website: https://www.bluenile.com/diamond-search
- Download HAR files for diamonds that are larger than **1.2** carat
- Keep the diamonds whose prices are lower than **$30,000**
- Data size: **9,985** unique diamonds with 10 different features



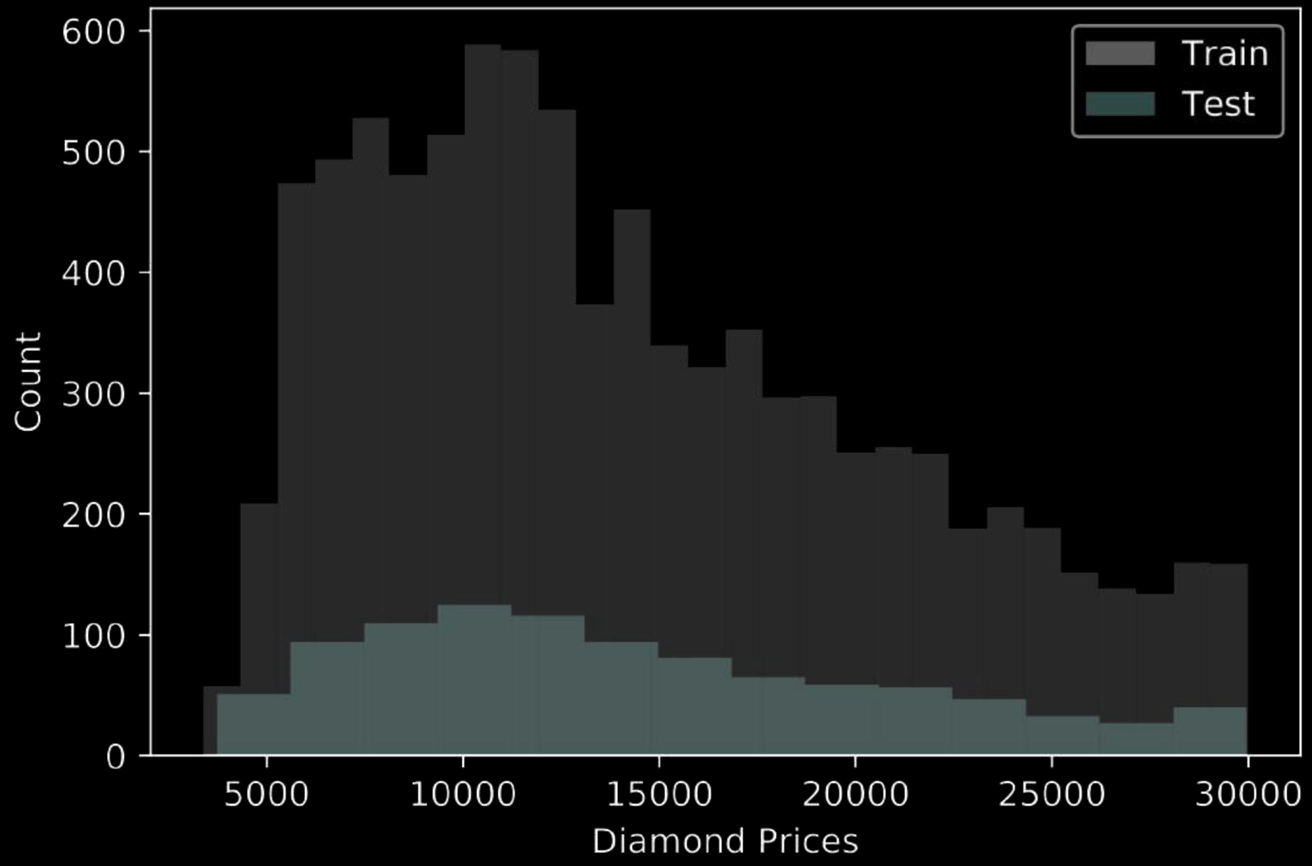
Feature	Numerical type
Carat	Float
Clarity	Category ([‘FL’, ‘IF’, ’VVS1’, ‘VVS2’, ‘VS1’, ‘VS2’, ‘SI1’, ‘SI2’])
Color	Category ([‘D’, ‘E’, ‘F’, ‘G’, ‘H’, ‘I’, ‘J’, ‘K’])
Culet	Category ([‘None’, ‘Pointed, ’Very Small, ‘Small’, ’Medium’, ’Slightly Large’ ,’Large’])
Cut	Category ([‘Astor Ideal’, ‘Ideal’, ‘Very Good’, ’Good’])
Depth	Float
Fluorescence	Category ([‘None’, ‘Faint’, ‘Faint Blue’, ‘Medium’, ‘Medium Blue’, ’Strong’, ‘Strong White’, ‘Strong Yellow’, ‘Strong Blue’, ‘Very Strong Blue’, ‘Very Strong’])
lxwRatio	Float
Polish	Category ([‘Excellent’, ‘Very Good’, ‘Good’])
Symmetry	Category ([‘Excellent’, ‘Very Good’, ‘Good’])
Table	Float

Data Statistics

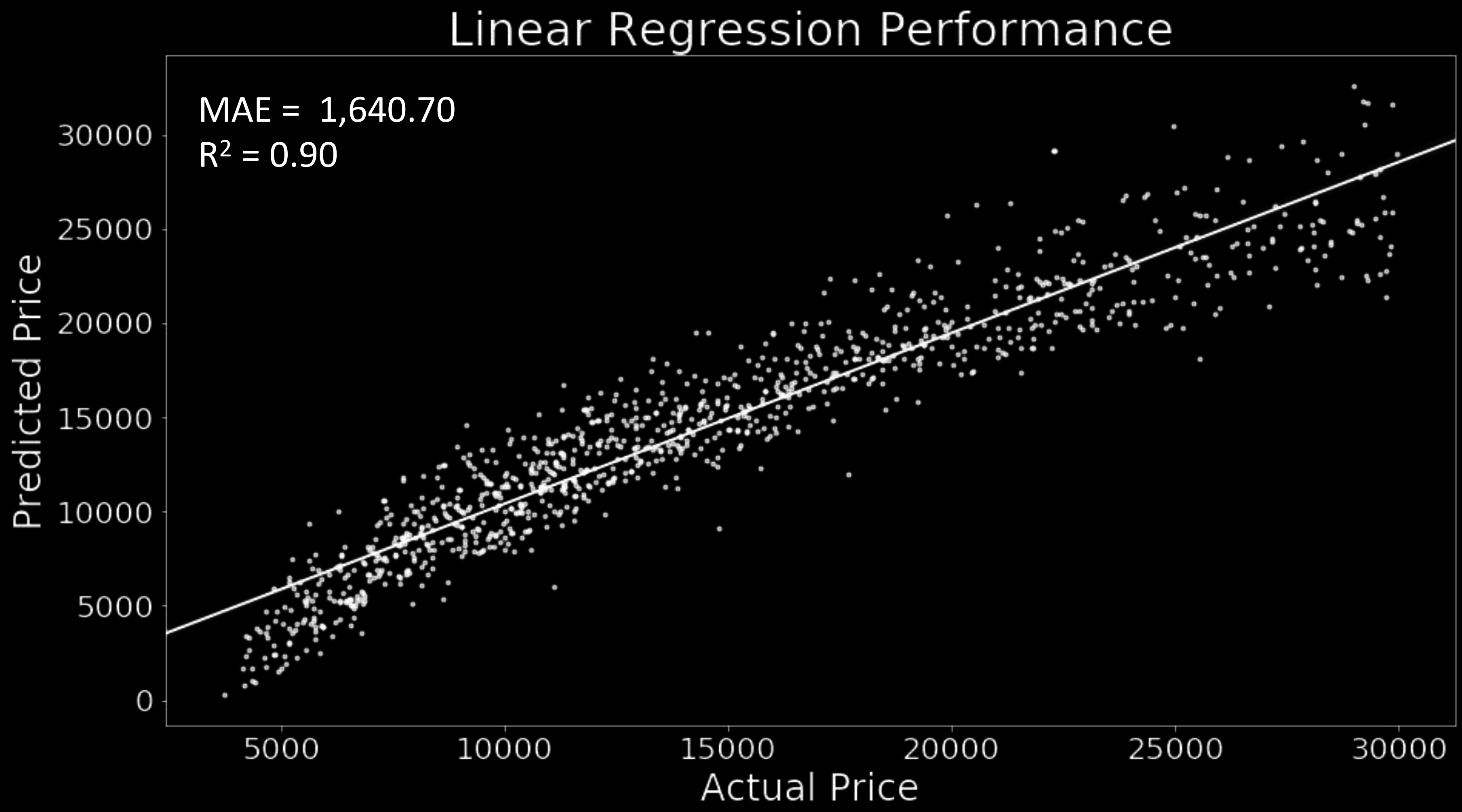
Label	Mean	Std	Min	Max
Price	14,387.39	6,592.27	3,375.00	29,995.00
Carat	1.56	0.33	1.20	3.06
Clarity	5.56	1.70	1	8
Color	4.54	2.08	1	8
Culet	1.09	0.43	1	5
Cut	2.12	0.37	2	4
Depth	62.16	1.20	55.30	71.20
Fluorescence	2.23	2.08	1	11
lxwRatio	1.00	0.004	1.00	1.03
Polish	1.05	0.25	1	3
Symmetry	1.13	0.37	1	3
Table	58.09	1.85	50.00	79.00

Train set & Test set

- Randomly split the data into train set and test set
- Train/Test size: 8,986/999

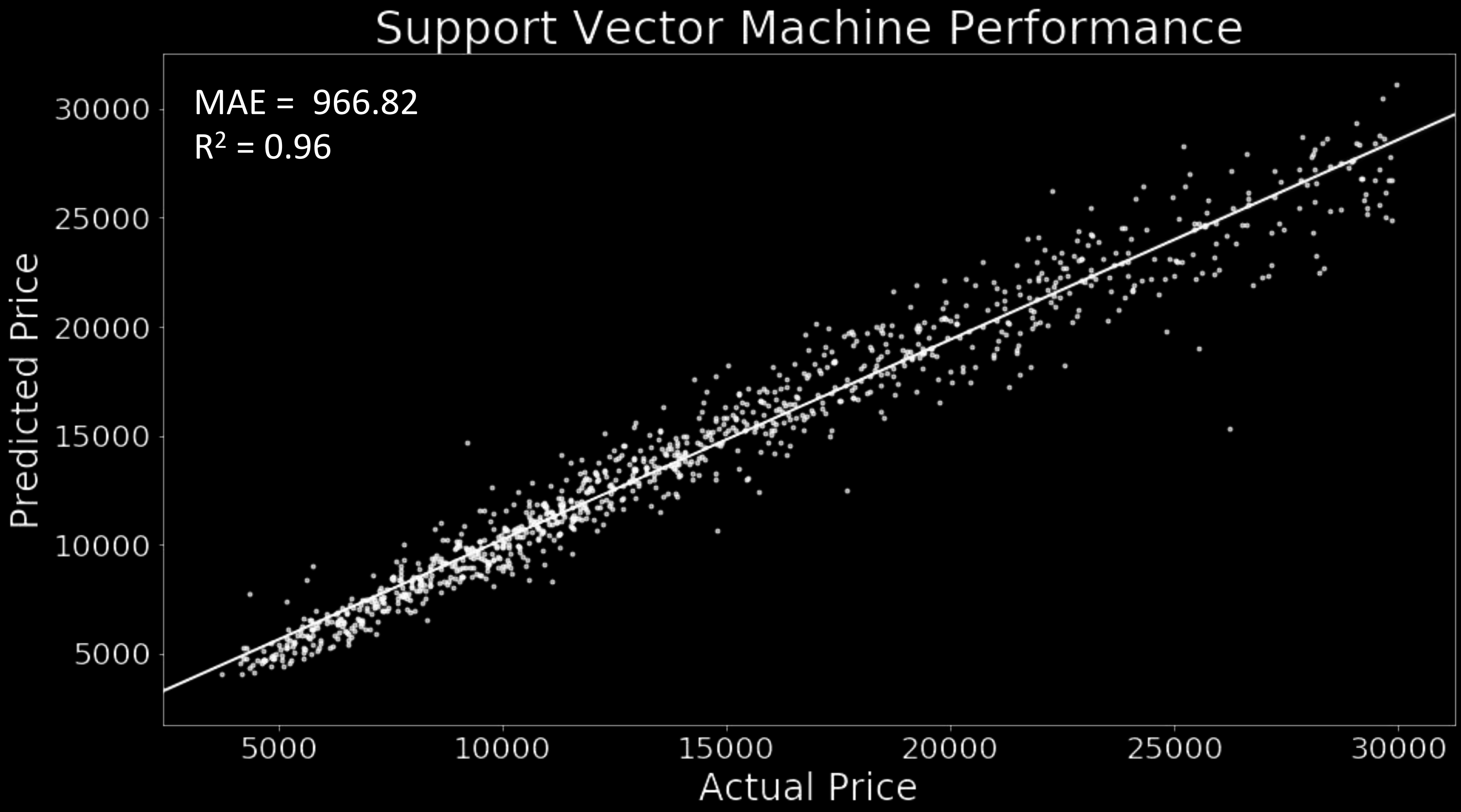


Linear Regression

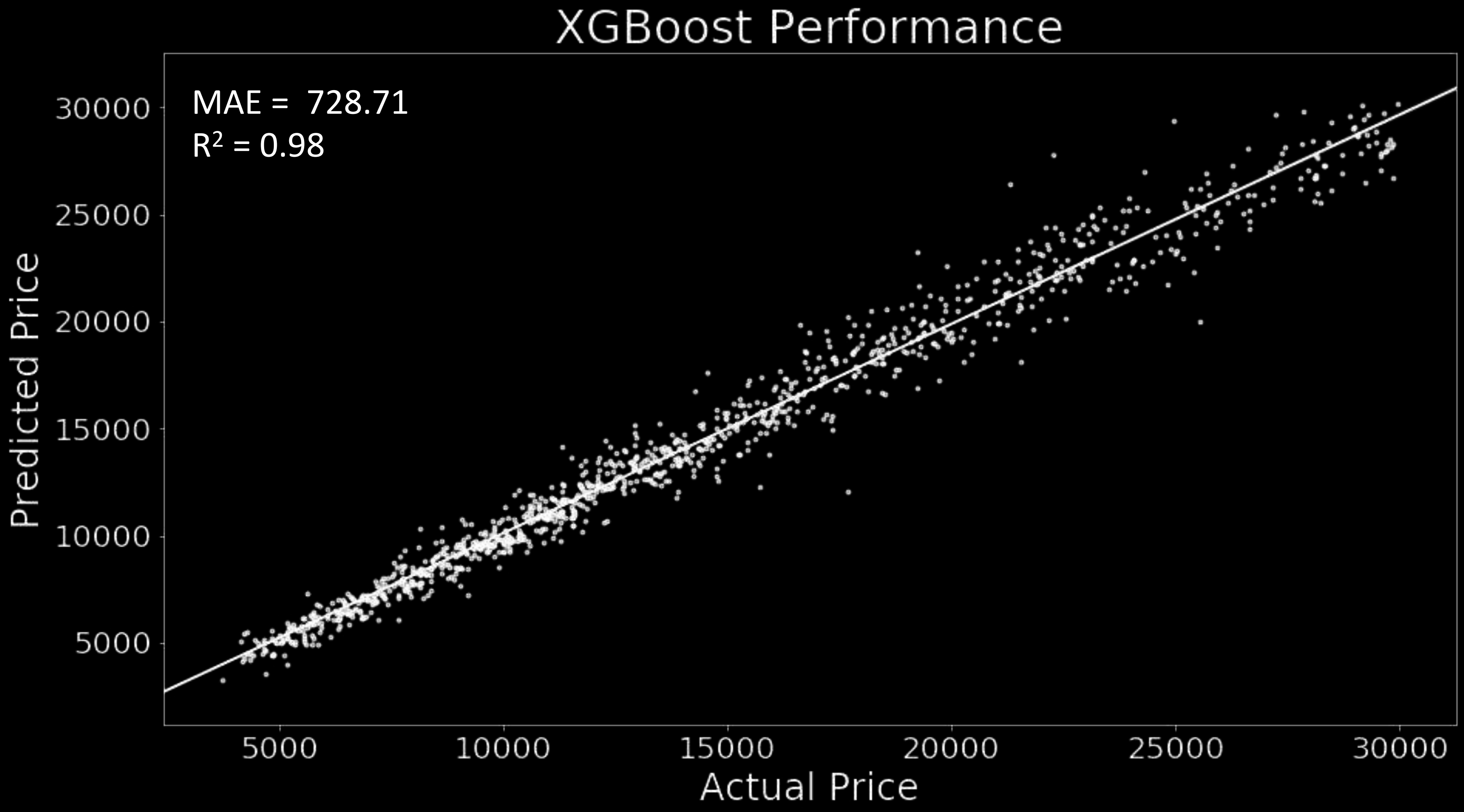

Linear Regression Performance

MAE = 1,640.70
R^2 = 0.90

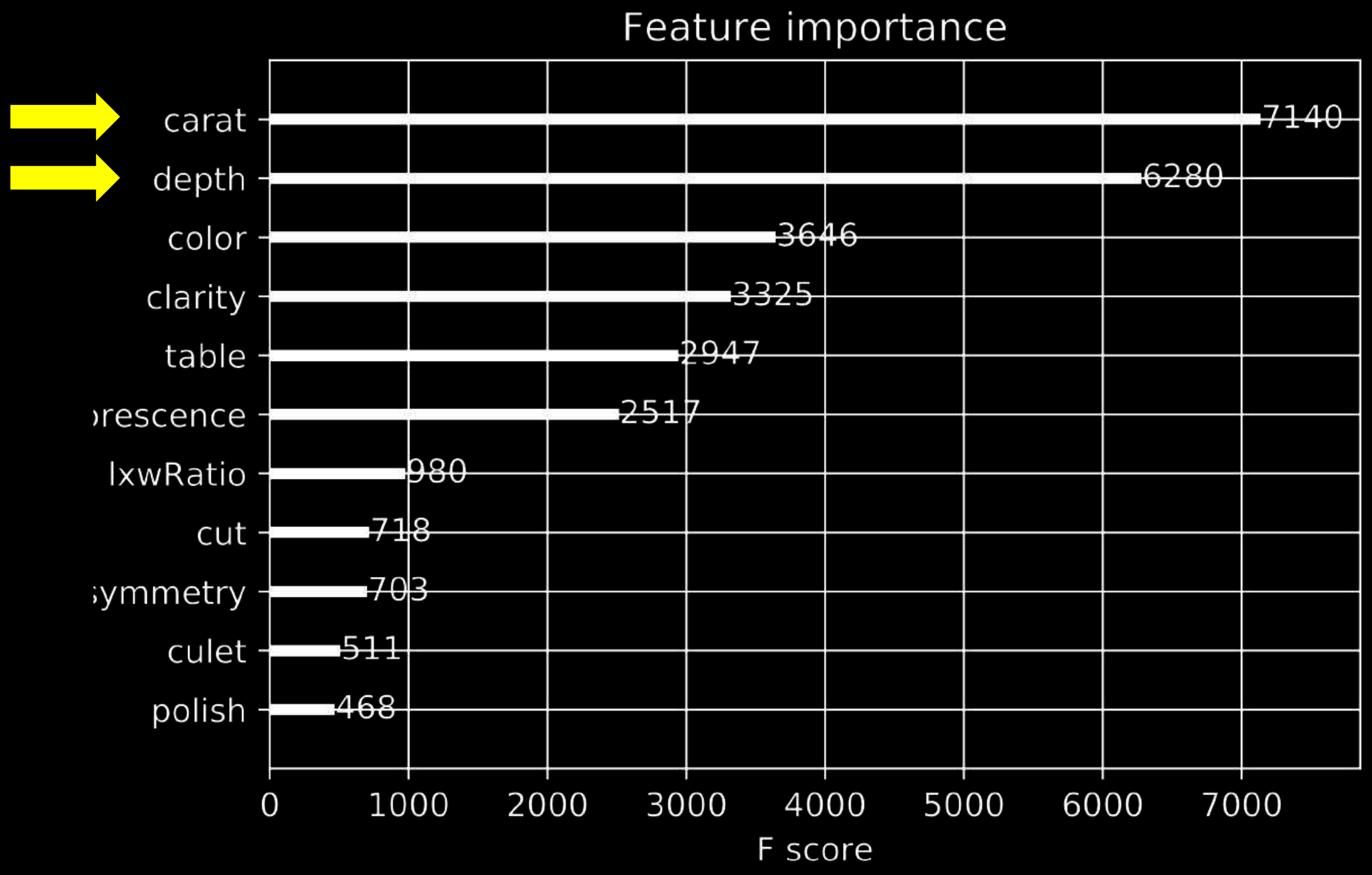
Support Vector Machine



XGBoost


XGBoost Performance
MAE = 728.71
R2 = 0.98
Predicted Price
30000
25000
20000
15000
10000
5000
5000
10000
15000
20000
25000
30000
Actual Price

XGBoost



Feature importance
carat
depth
color
clarity
table
rescence
lxwRatio
cut
symmetry
culet
polish
7140
6280
3646
3325
2947
2517
980
718
703
511
468
0
1000
2000
3000
4000
5000
6000
7000
F score

Thank you