Shine Bright Like A Diamond

Building Model to Predict Diamond Price

Presented by Shuo Jia

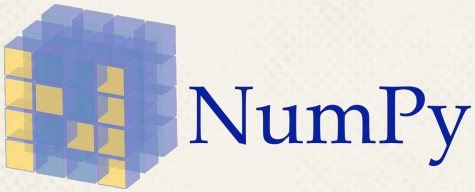
Objective

With continued economic growth around the world and more demand for luxury goods and decline in the mining of diamonds, it is expected that the prices of high gem quality natural diamonds will increase by roughly 6% each year.

- **To predict price when given the measurement parameters of loose diamonds**
- **To Encourage Rational Consumption and investment**

Data & Tools

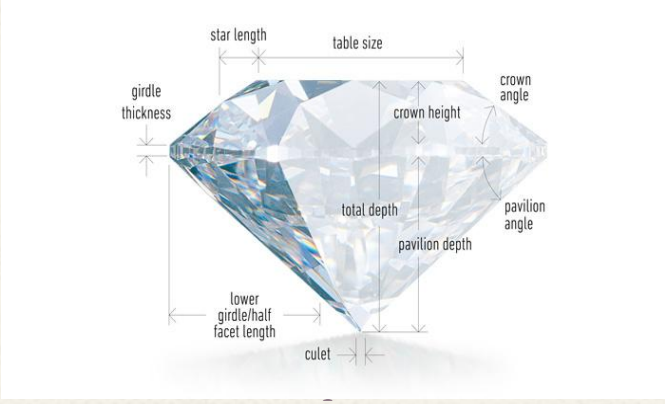
- **Data source: Diamond Search Engine**
- **Reference:** *Understanding Diamond Depth And Table Percentages*



Understanding Data

- Cut - Good, Very Good, Ideal
- Clarity - I2, I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF, FL
- Color - L, K, J, I, H, G, F, E, D
- Carat - The weight of diamond
- Length/Width/Depth: measurements in mm
- Diamond Shape - Round, oval, heart, asscher, emerald, marquise, pear, cushion, princess, radiant

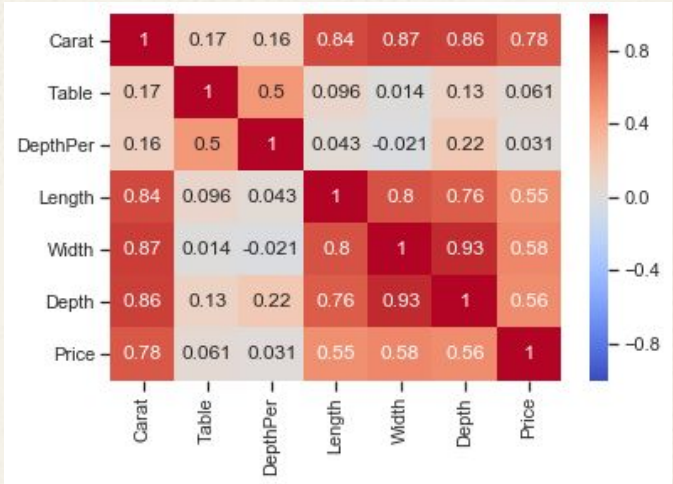
- **Table%** - Table width divided by avg girdle diameter of the diamond
- **Depth%** - The height of diamond divided by avg girdle diameter



Process Workflow

- Web scraping - more than 360,000 rows of data
- Data cleaning & EDA
- Feature selection & engineering:
  - Log transform Price
  - Run Variance Inflation Factor
  - New feature: table% * depth%
  - Drop feature with high VIF score

	VIF Factor	features
0	9.674879	Carat
1	262.722278	Table
2	305.550569	DepthPer
3	51.460793	Length
4	164.627158	Width
5	174.345288	Depth



Model Training & Validation

- Features - all the original predictions without transformation
- Predicted value - y log transformed
- Model methods
  - Linear regression
  - Polynomial regression with degree of 2
  - Ridge regression
- Cross validation - split into 10 folds -

Cross Validation Results

	Adjusted R Squared	Standard Deviation
Linear Regression	0.96	+-0.001
Ridge Regression	0.96	+-0.001
Polynomial Regression	-0.981	+-0.006

Testing Result

- **BAM!!!**

- **Adjusted R squareds are extremely low…**
  - **Ridge: -0.0702**
  - **Linear regression: -0.0705**
  - **Polynomial: -0.0702**

What Does It Mean?

```
[('Carat', -0.28228491239035935),
 ('Table', 0.06413204294365665),
 ('DepthPer', 0.21987355869113653),
 ('Length', 0.4814504654023417),
 ('Width', 0.9728026015616641),
 ('Depth', 0.054717755527994134),
 ('Price', 0.005964643424203499),
 ('Shape_fancy', -0.005964643423005309),
 ('Shape_round', -0.022639514161563616),
 ('Cut_Good', 0.00823579955429188),
 ('Cut_Ideal', 0.005214550203678386),
 ('Cut_V.Good', 0.056225023796657725),
 ('Color_D', 0.03956481360469803),
 ('Color_E', 0.03212259065545117),
 ('Color_F', 0.01350340811106751),
 ('Color_G', -0.0109912546082638),
 ('Color_H', -0.04637283028098869),
 ('Color_I', -0.071782073220157),
 ('Color_J', -0.07425004834854336),
 ('Color_K', -0.06584818494543829),
 ('Color_L', 0.03598942813539178),
 ('Clarity_FL', -0.06861794139539419),
 ('Clarity_I1', 0.054280879499801975),
 ('Clarity_IF', -0.044700513925241186),
 ('Clarity_SI1', -0.08742196127279332),
 ('Clarity_SI2', 0.024194263109531 42),
 ('Clarity_VS1', 0.002291323937364203),
 ('Clarity_VS2', 0.058254609440117365),
 ('Clarity_VVS1', 0.036485648997086 04)]
```

- **Coefficient of Ridge model**
- **Seems odd…**
- **Multicollinearity..**

Future Work

- **Definitely work on Multicollinearity!**
- **Try principal component analysis**

THANKS!

Any questions?