Diamond

SOPHISTICATED

Toronto | Vancouver | Quebec

Loose Diamond

Price Prediction

Diamond
SOPHISTICATED

Agenda

What we'll discuss today

01

Introduction

02

Business Challenges & Impacts

03

Data Pre-Processing &
Data Pattern Visualizing in Dataset

04

Machine Learning Techniques &
Outlook highlighted

05

Conclusion

Diamond
SOPHISTICATED

Introduction

Product: Loose Diamonds

Services:

1) Offer an exclusive selection of loose diamonds and discover the ideal gem to complement the customer's jewellery, considering factors such as clarity, cut, carat, and size.

2) Facilitate the purchase of loose diamonds from customers, ensuring a professional and seamless process with a reasonable price for those looking to sell their precious ones.

Diamond Sophisticated's Vision:

- Enhance sales conversion rates
- Acquire unique diamonds at competitive prices
- Avoid overestimation or underestimation
- Implement a robust diamond pricing model

Diamond
SOPHISTICATED

Business Challenges

Sales representatives estimate product price using their professional experience

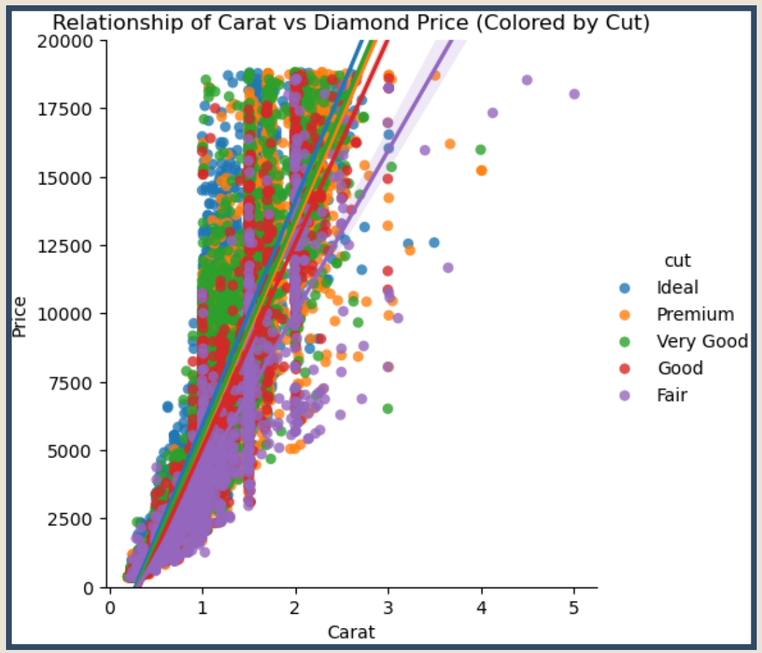
Customers' hesitation or seeking an alternative way for purchases or sales because of inconsistency diamond prices

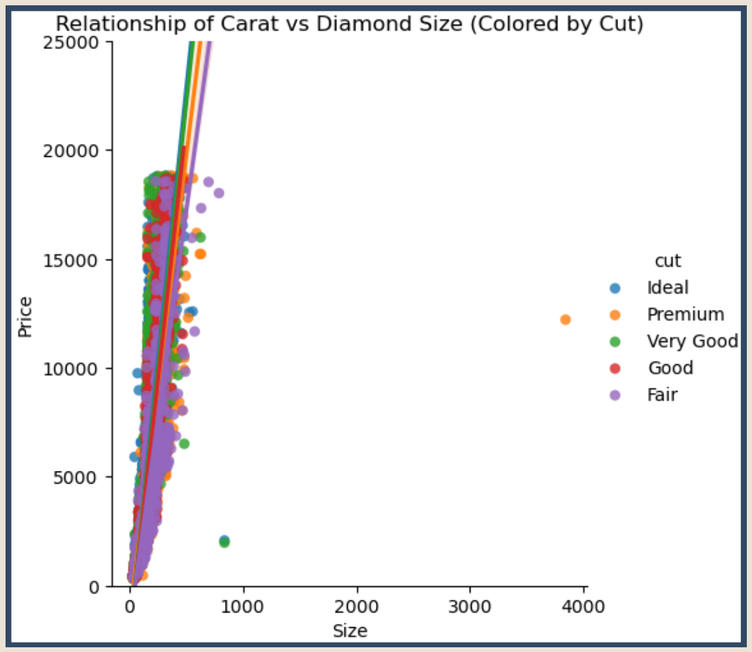
Business Impacts

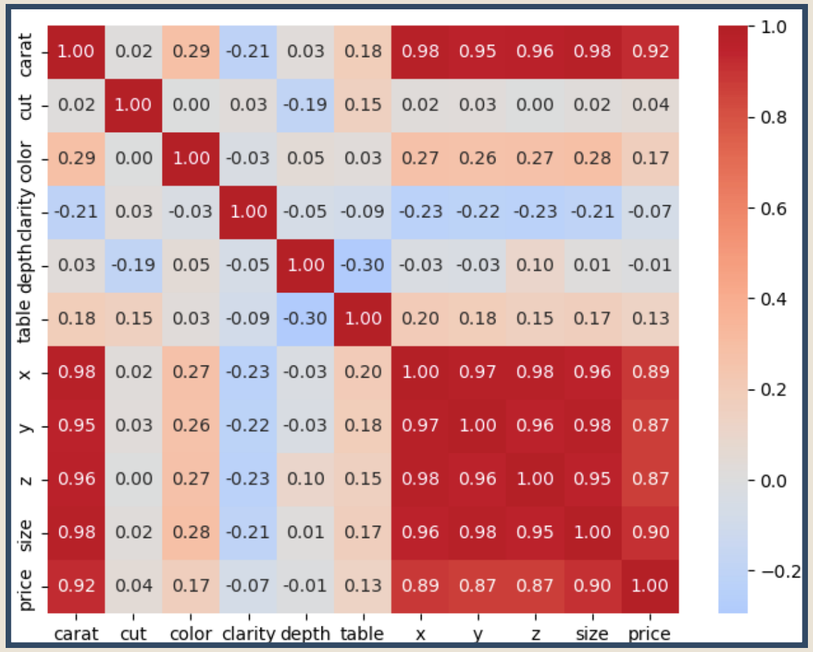
Misjudging diamond prices, whether overestimating or underestimating, has shown a direct impact, causing a 35% reduction in sales conversion rates and a notable 25% decline in the acquisition of unique diamonds from customers.

Diamond
SOPHISTICATED

Data Pattern Visualizing in Dataset:







Data-Preprocessing: Remove dimensionless diamonds & Create new variable “size”, which is calculated by x*y*z.

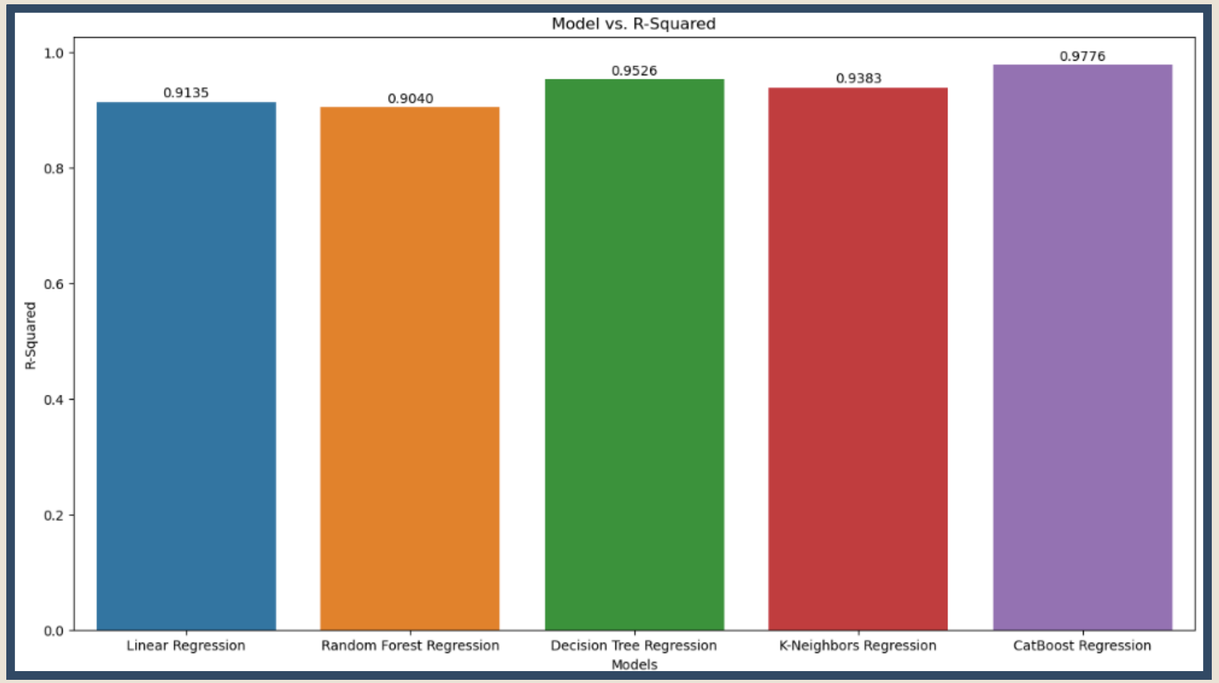
Data Pattern Highlights:

1) The carat vs price has positive relationship, especially ideal cut.

2) It seems like the size vs price is positive relationship, no matter what the cut is.

3) Referring to correlation heatmap, the size vs price is positive correlation with 0.90. It can be explained that if the size is bigger, the price is higher.

ML Techniques & Outlook highlighted:



- Linear Regression
- Random Forest Regression
- Decision Tress Regression
- K-Neighbors Regression
- CatBoost Regression

The reasons why we chose 5 ML Technique are:

1) There are categorical variables (like diamond cut, colour, and clarity) and have a relatively linear relationship with certain features like size vs price. With those characteristics, CatBoost & Linear Regression are suitable for building the model.

2) Both CatBoost and linear regression, provide interpretability and good performance.

3) The algorithm of KNN works well when the data points with similar features tend to have similar prices. It might capture this effectively.

Highlights:

CatBoost Regression stands out as a top-performing model with the highest R-squared (= 0.98), but it can be more computationally expensive during training compared to simpler models like Linear Regression.

Diamond
SOPHISTICATED

Conclusion

Model Name	R-Squared	MSE	MAE	MAPE	Cross Validation R2
Linear Regression	0.91	1314185.00	731.6	39.75%	0.84
Random Forest Regression	0.90	1457220.9	639.9	15.95%	0.89
Decision Tree Regression	0.95	719436.60	439.4	11.47%	0.94
K-Neighbors Regression	0.94	936348.10	500.7	12.87%	0.93
CatBoost Regression	0.98	340096.20	299.8	8.87%	0.98

1. CatBoost stands out as the superior choice.
2. High R-square values, lower errors (MAE, RMSE), and superior cross-validation results.
3. In scenarios where the primary focus is on optimizing predictive performance, CatBoost emerges as the best choice.