# Introduction

To predict 2024 diamond prices, we developed several predictive models. The first models were built on data that was trained on 70% of the 10 year diamond sales. Because the dataset contained outliers which could cause errors in our models, a meta-model approach was taken. The dataset was clustered into 2 sets, Outliers and Normal Data. The models were trained on 70% on outlier data and 70% on normal data respectively and validated using the remaining 30%. If all the company’s diamonds are sold in 2024, the total sales will be as indicated in Table 1, based on the selected model. In Table 2, we have mapped out the total sales per cut to get an idea on which cut is the most profitable and should be focused on.

### Table 1: Total Predicted Sales and Average Sale Price

| **Model** | **Prediction state** | **Total sales** | **Average diamond price** |
| --- | --- | --- | --- |
| Random Forest | Without Hyperparameter tuning | $139,892,669 | $4,121 |
|  | Based on outlier trained cluster | $390,947,200 | $11,518 |
|  | Based on cluster without outliers | $139,892,669 | $4,121 |
|  | Metamodel (combined prediction based on two clusters) | $265,419,934 | $7,820 |
| LGBM | Without Hyperparameter tuning | $139,063,307 | $4,097 |
|  | Based on outlier trained cluster | $388,653,820 | $11,451 |
|  | Based on cluster without outliers | $139,063,307 | $4,097 |
|  | Metamodel (combined prediction based on two clusters) | $263,858,563 | $7,774 |
| XGB | Without Hyperparameter tuning | $138,996,704 | $4,095 |
|  | Based on outlier trained cluster | $393,830,048 | $11,603 |
|  | Based on cluster without outliers | $138,996,704 | $4,095 |
|  | Metamodel (combined prediction based on two clusters) | $266,413,376 | $7,849 |
| KNN | Without Hyperparameter tuning | $137,478,534 | $4,050 |
|  | Based on outlier trained cluster | $465,417,758 | $13,712 |
|  | Based on cluster without outliers | $137,478,534 | $4,050 |
|  | Metamodel (combined prediction based on two clusters) | $301,448,146 | $8,881 |
| Decision Tree | Without Hyperparameter tuning | $140,015,783 | $4,125 |
|  | Based on outlier trained cluster | $389,640,043 | $11,480 |
|  | Based on cluster without outliers | $140,181,432 | $4,130 |
|  | Metamodel (combined prediction based on two clusters) | $264,910,738 | $7,805 |
| Linear Regression | Without Hyperparameter tuning | $143,733,580 | $4,234 |
|  | Based on outlier trained cluster | $487,811,444 | $14,372 |
|  | Based on cluster without outliers | $143,815,339 | $4,237 |
|  | Metamodel (combined prediction based on two clusters) | $315,813,392 | $9,305 |

### Table 2: Predicted Sales for Each Cut Type

| **Model** | **Prediction state** | **Sales for each Cut Type** | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Fair** | **Good** | **Ideal** | **Premium** | **Very good** |
| Random Forest | Without Hyperparameter tuning | $4,683,530 | $12,907,915 | $49,281,704 | $40,969,363 | $32,050,155 |
|  | Based on outlier trained cluster | $12,402,131 | $36,537,731 | $148,141,666 | $104,262,180 | $89,603,489 |
|  | Based on cluster without outliers | $4,683,530 | $12,907,915 | $49,281,704 | $40,969,363 | $32,050,155 |
|  | Metamodel (combined prediction based on two clusters) | $8,542,830 | $24,722,823 | $98,711,685 | $72,615,772 | $60,826,822 |
| LGBM | Without Hyperparameter tuning | $4,599,103 | $12,862,761 | $49,028,546 | $40,645,248 | $31,927,647 |
|  | Based on outlier trained cluster | $12,346,799 | $36,852,956 | $146,316,522 | $104,012,108 | $89,125,433 |
|  | Based on cluster without outliers | $4,599,103 | $12,862,761 | $49,028,546 | $40,645,248 | $31,927,647 |
|  | Metamodel (combined prediction based on two clusters) | $8,472,951 | $24,857,858 | $97,672,534 | $72,328,678 | $60,526,540 |
| XGB | Without Hyperparameter tuning | $4,580,829 | $12,857,750 | $48,973,380 | $40,662,768 | $31,921,984 |
|  | Based on outlier trained cluster | $12,412,139 | $37,472,608 | $148,127,024 | $105,277,800 | $90,540,488 |
|  | Based on cluster without outliers | $4,580,829 | $12,857,750 | $48,973,380 | $40,662,768 | $31,921,984 |
|  | Metamodel (combined prediction based on two clusters) | $8,496,484 | $25,165,178 | $98,550,200 | $72,970,280 | $61,231,236 |
| KNN | Without Hyperparameter tuning | $4,478,312 | $12,409,551 | $48,707,819 | $40,319,165 | $31,563,684 |
|  | Based on outlier trained cluster | $9,948,613 | $34,863,842 | $2,068,077,836 | $112,493,276 | $101,304,190 |
|  | Based on cluster without outliers | $4,478,312 | $12,409,551 | $48,707,819 | $40,319,165 | $31,563,684 |
|  | Metamodel (combined prediction based on two clusters) | $7,213,462 | $23,636,697 | $127,757,828 | $76,406,220 | $66,433,937 |
| Decision Tree | Without Hyperparameter tuning | $4,654,058 | $12,919,815 | $49,341,901 | $49,341,901 | $32,104,470 |
|  | Based on outlier trained cluster | $12,505,231 | $36,878,553 | $147,191,947 | $104,267,080 | $88,797,230 |
|  | Based on cluster without outliers | $4,700,033 | $12,919,008 | $49,398,465 | $41,032,096 | $32,131,828 |
|  | Metamodel (combined prediction based on two clusters) | $8,602,632 | $24,898,780 | $98,295,206 | $72,649,588 | $60,464,529 |
| Linear Regression | Without Hyperparameter tuning | $5,137,585 | $13,331,399 | $49,878,808 | $42,602,105 | $32,783,681 |
|  | Based on outlier trained cluster | $14,749,815 | $41,087,048 | $191,284,551 | $130,419,037 | $110,270,992 |
|  | Based on cluster without outliers | $51,370 | $13,341,313 | $49,960,036 | $42,637,279 | $32,739,687 |
|  | Metamodel (combined prediction based on two clusters) | $9,943,418 | $27,214,180 | $120,622,294 | $86,528,158 | $71,505,340 |

# Conclusion

Our predictive analysis for 2024 diamond sales highlights significant insights:  
The metamodel approach projects an optimal blend of accuracy and insight, with Random Forest and XGBoost models indicating a notable increase in average prices and total sales. The analysis suggests a strong market preference for Ideal and Premium cuts, pointing towards higher value transactions. We suggest the company should focus on these cuts in order to capitalize on potential sales growth in 2024.

In the future, we would like to look at price variation based on year to get an idea on if 2024 would be better performing compared to previous years. This could give the company insight on if there would need to be some type of measures implemented to boost sales and stay on track.