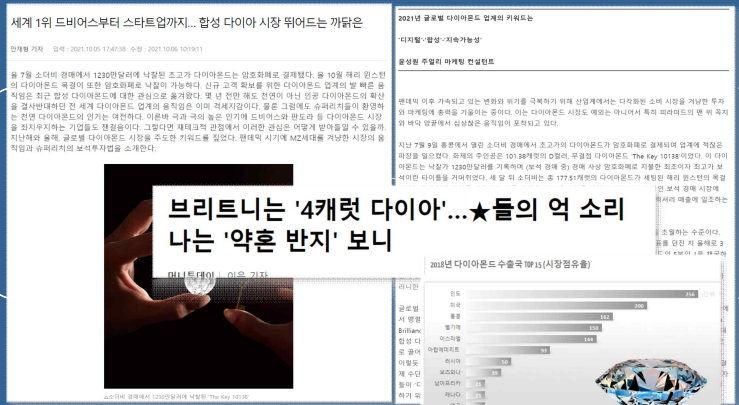
다이아몬드 가격 예측 프로젝트

보석의 제왕 '다이아몬드'

조아진

프로젝트 소개 및 기획 배경


세계 1위 드비어스부터 스타트업까지... 합성 다이아 시장 뒤어드는 까닭은

안재형 기자 | 입력 : 2021.10.05 17:47:38 수정 : 2021.10.06 10:19:11

올 7월 소더비 경매에서 1230만달러에 낙찰된 조고가 다이아몬드는 암호화폐로 결제됐다. 올 10월 래리 윈스턴의 다이아몬드 목걸이 또한 암호화폐로 낙찰이 가능하다. 신규 고객 확보를 위한 다이아몬드 업계의 발 빠른 움직임은 최근 합성 다이아몬드에 대한 관심으로 옮겨왔다. 몇 년 전만 해도 천연이 아닌 인공 다이아몬드의 확산을 걸사반대하던 전 세계 다이아몬드 업계의 움직임은 이제 격세지감이다. 물론 그럼에도 슈퍼리치들이 환영하는 한편 다이아몬드의 인기는 여전하다. 이른바 국과 국의 높은 인기를 드비어스와 판도라 등 다이아몬드 시장을 좌지우지하는 기업들도 잡질을이다. 그렇다면 재테크적 관점에서 이러한 관심은 어떻게 받아들일 수 있을까. 지난해와 올해, 글로벌 다이아몬드 시장을 주도한 키워드를 짚었다. 팬데믹 시기에 MZ세대를 겨냥한 시장의 움직임과 슈퍼리치의 보석투자법을 소개한다.

브리트니는 '4캐럿 다이아'...★들의 억 소리 나는 '약혼 반지' 보니

머니투데이 이은 기자

소더비에 경매에서 1230만달러에 낙찰된 'The Key 10138'

2021년 글로벌 다이아몬드 업계의 키워드는

'디지털'·'합성'·'지속가능성'

윤성원 주얼리 마케팅 컨설턴트

팬데믹 이후 가속되고 있는 변화와 위기를 극복하기 위해 산업계에서는 다각화된 소비 시장을 겨냥한 투자와 마케팅에 총력을 기울이는 중이다. 이는 다이아몬드 시장도 예외는 아니어서 특히 피라미드의 맨 위 꼭지와 바닥 양끝에서 심상찮은 움직임이 포착되고 있다.

지난 7월 9일 홍콩에서 열린 소더비 경매에서 조고가의 다이아몬드가 암호화폐로 결제되며 업계에 적잖은 파장을 일으켰다. 화제의 주인공은 101.38캐럿의 D컬러, 무결점 다이아몬드 'The Key 10138'이었다. 이 다이아몬드는 낙찰가 1230만달러를 기록하며 (보석 경매 중) 경매 사상 암호화폐로 지불한 최초이자 최고가 보석이란 타이틀을 거머쥐었다. 세 달 뒤 소더비는 총 177.51캐럿의 다이아몬드가 세팅된 래리 윈스턴의 목걸이 보석 경매 시장에서 매출액 알조하는 초월하는 수준이다. 로 인한 지 올해로 3

2018년 다이아몬드 수출국 TOP 15 (시장점유율)

다이아몬드 가격 예측 모델을 활용해 다이아몬드 규격에 알맞는 적정 가격 정보를 제공하여 판매자와 소비자간 정보 비대칭을 완화하는데 목적이 있음.

목차

1 step

- 데이터 전처리
- 데이터 분석
- 시각화 EDA

2 step

- 모델링
- 최적 모델 선정

3 step

- 서비스 구현
- 기대효과(결론)

데이터 전처리

1. 결측치(Null) 확인

```
diaDF=pd.read_csv('data/diamonds.csv', encoding='cp949')
diaDF
```

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...	...
53935	53936	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	53937	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	53938	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	53939	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	53940	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 11 columns

```
#결측치 확인
diaDF.isna().sum()
```

```
Unnamed: 0    0
carat         0
cut           0
color         0
clarity       0
depth         0
table         0
price         0
x             0
y             0
z             0
dtype: int64
```

2. 불필요 칼럼 제거 및 기초통계량 값 확인

```
#Unnamed행은 단순 숫번나열이므로 삭제
diaDF=diaDF.drop(['Unnamed: 0'],axis=1)
diaDF
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53940 rows × 10 columns

```
#53940개 데이터, 10개 변수
diaDF.shape
```

```
(53940, 10)
```

```
#기초통계량 값 확인
diaDF.describe()
```

3. 결측치 처리 및 Dtype 매칭

```
# x(길이), y(너비), z(깊이)가 0인 값은 결측치로 간주하여 삭제
diaDF.drop(diaDF[diaDF['x']==0].index,inplace=True)
diaDF.drop(diaDF[diaDF['y']==0].index,inplace=True)
diaDF.drop(diaDF[diaDF['z']==0].index,inplace=True)
```

```
#기초통계량
diaDF.describe()
```

	carat	depth	table	price	x	y	z
count	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000	53920.000000
mean	0.797698	61.749514	57.456834	3930.993231	5.731627	5.734887	3.540046
std	0.473795	1.432331	2.234064	3987.280446	1.119423	1.140126	0.702530
min	0.200000	43.000000	43.000000	326.000000	3.730000	3.680000	1.070000
25%	0.400000	61.000000	56.000000	949.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5323.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

```
#Dtype 매칭 확인
diaDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 53920 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53920 non-null  float64
 1   cut      53920 non-null  object
 2   color    53920 non-null  object
 3   clarity  53920 non-null  object
 4   depth    53920 non-null  float64
 5   table    53920 non-null  float64
 6   price    53920 non-null  int64
 7   x        53920 non-null  float64
 8   y        53920 non-null  float64
 9   z        53920 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.5+ MB
```

데이터 소개

전처리 후 data

diaDF

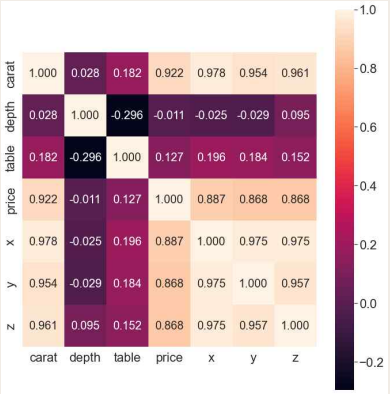
	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53935	0.72	Ideal	D	SI1	60.8	57.0	2757	5.75	5.76	3.50
53936	0.72	Good	D	SI1	63.1	55.0	2757	5.69	5.75	3.61
53937	0.70	Very Good	D	SI1	62.8	60.0	2757	5.66	5.68	3.56
53938	0.86	Premium	H	SI2	61.0	58.0	2757	6.15	6.12	3.74
53939	0.75	Ideal	D	SI2	62.2	55.0	2757	5.83	5.87	3.64

53920 rows × 10 columns

변수 설명

변수명	값 설명
carat	다이아몬드의 캐럿 무게
cut	다이아몬드의 절단 품질 (Fair, Good, Very Good, Premium, Ideal
color	다이아몬드 색상, J(최악)에서 D(최상)
clarity	다이아몬드의 투명도 측정(I1(최악), SI2, SI1, VS2, VS1, VVS2, VVS1, IF(최고))
depth	큐렛에서 테이블까지 측정한 다이아몬드의 높이를 평균 거들 직경으로 나눈 값(%)
table	다이아몬드 테이블의 너비는 평균 직경의 백분율로 표시
price	다이아몬드의 가격
x	길이(mm)(0--10.74)
y	너비(mm)(0--58.9)
z	깊이(mm(0--31.8)

변수 별 상관분석(heatmap)



- 무게 & 가격 : 0.92 높은 양의 상관관계
- 무게 & 길이 : 0.97 높은 양의 상관관계
- 무게 & 너비 : 0.95 높은 양의 상관관계
- 무게 & 깊이 : 0.96 높은 양의 상관관계

- 가격 & 무게 : 0.92 높은 양의 상관관계
- 가격 & 길이 : 0.88 높은 양의 상관관계
- 가격 & 너비 : 0.86 높은 양의 상관관계
- 가격 & 깊이 : 0.86 높은 양의 상관관계

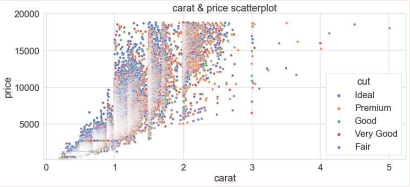
→ 다이아몬드의 무게가 클수록 가격이 높고, 길이,너비,깊이의 크기가 클수록 가격이 높아지는 양의 상관관계가 있음.

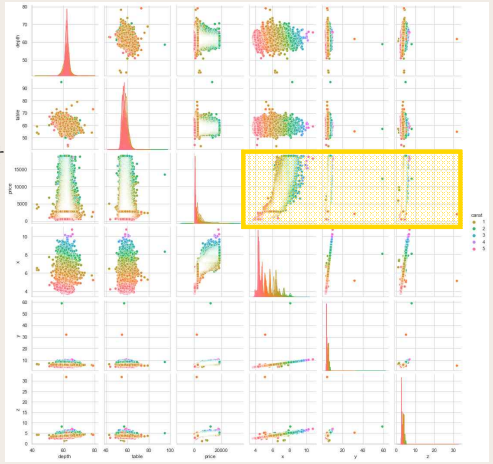
시각화 EDA

변수별 히스토그램과 두 변수 사이의 scatter plot을 한 번에 보여주는 pairplot

- depth와 table 변수를 제외한 연속변수 (price, x, y, z)들은 변수간 양의 상관관계 나타남

- carat의 무게, x(길이), y(너비), z(깊이)는 다이아몬드의 가격에 크게 영향을 주는 변수임 (조금만 상승해도 가격이 가파르게 상승하는 형태)

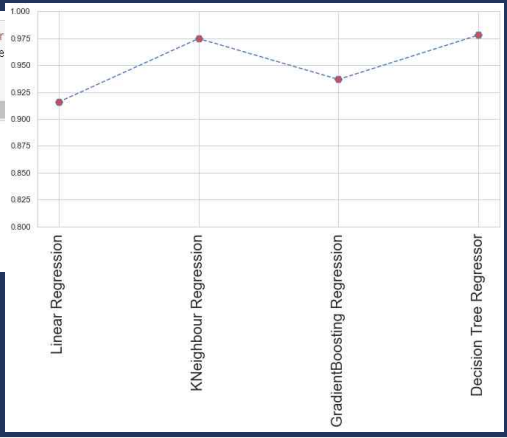




예측모델 생성

```
models=pd.DataFrame({'Model':['Linear Regression','KNeighbour
                     'Score':[r2_lr,r2_kneigh,r2_gbr,r2_dtree
                    })
models
```

	Model	Score
0	Linear Regression	0.915695
1	KNeighbour Regression	0.974651
2	GradientBoosting Regression	0.936713
3	Decision Tree Regressor	0.977872



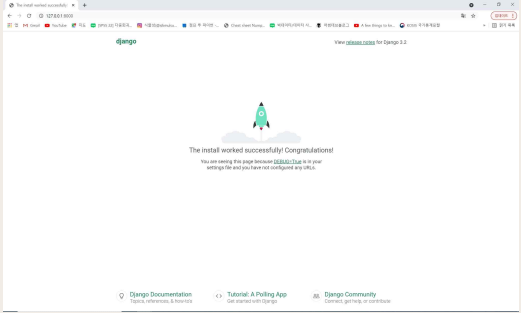
최적모델선정

알고리즘	accuracy_score	예측값
Linear Regression	0.91	price
KNeighbours Regressor	0.97	price
Decision Tree Regressor	0.97	price
Gradient Boosting Regressor	0.93	price

서비스구현

서비스 내용

- 다이아몬드의 무게, 길이, 너비, 깊이 등을 입력하면 **적정 가격 제시**
- 판매자와 소비자 양측에 해당 다이아몬드의 적정 가격 정보 제공하여 **정보 비대칭 완화**



감사합니다