Machine Learning을 활용한
Melbourne Housing Market Data 분석

데이터인텔리전스 팀
정상교 송하규 박범선

2017. 09. 08

LG CNS

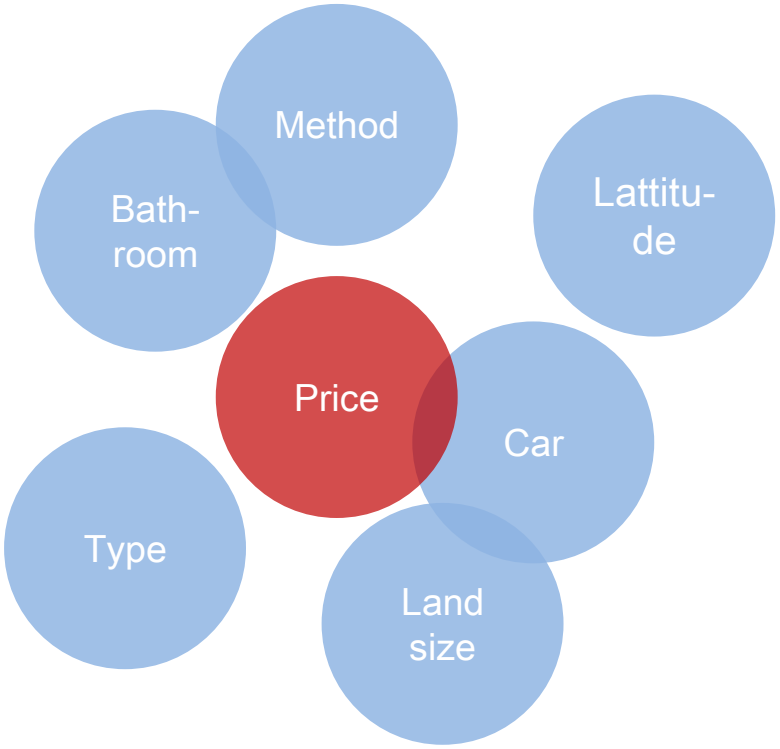
1. 분석 목적

Melbourne Housing Market 데이터를 Data Mining 및 Machine Learning 기법을 활용하여 Price를 예측하는 모델을 만들어 분석함

1. 데이터 전처리 및 EDA

2. 모델 적용 및 분석
   1) Random Forest
   2) Support Vector Machine
   3) Decision Tree + Regression

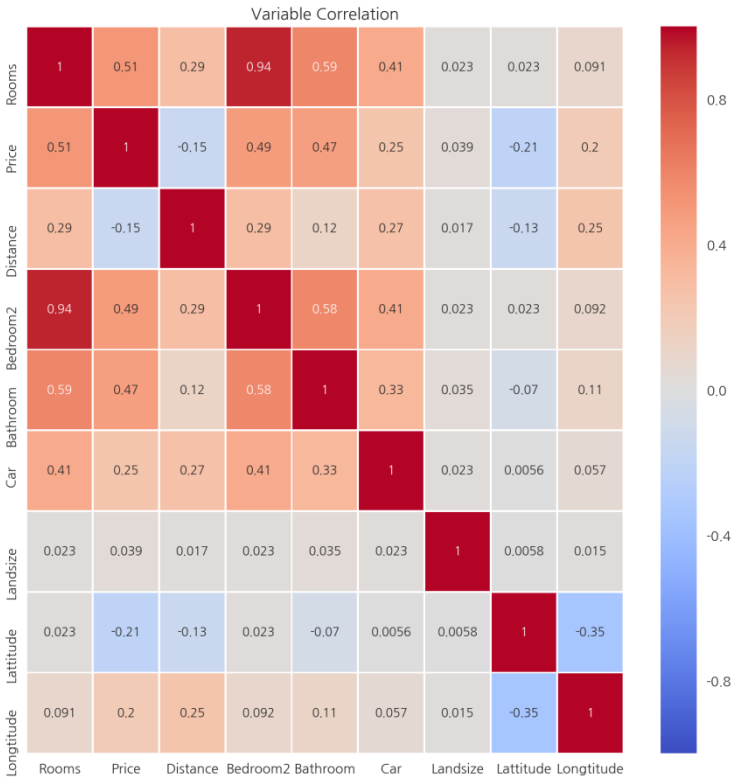
3. Summary



LG CNS

2. 데이터 전처리 및 EDA

변수간의 상관관계를 분석하고 데이터 마트를 완성함



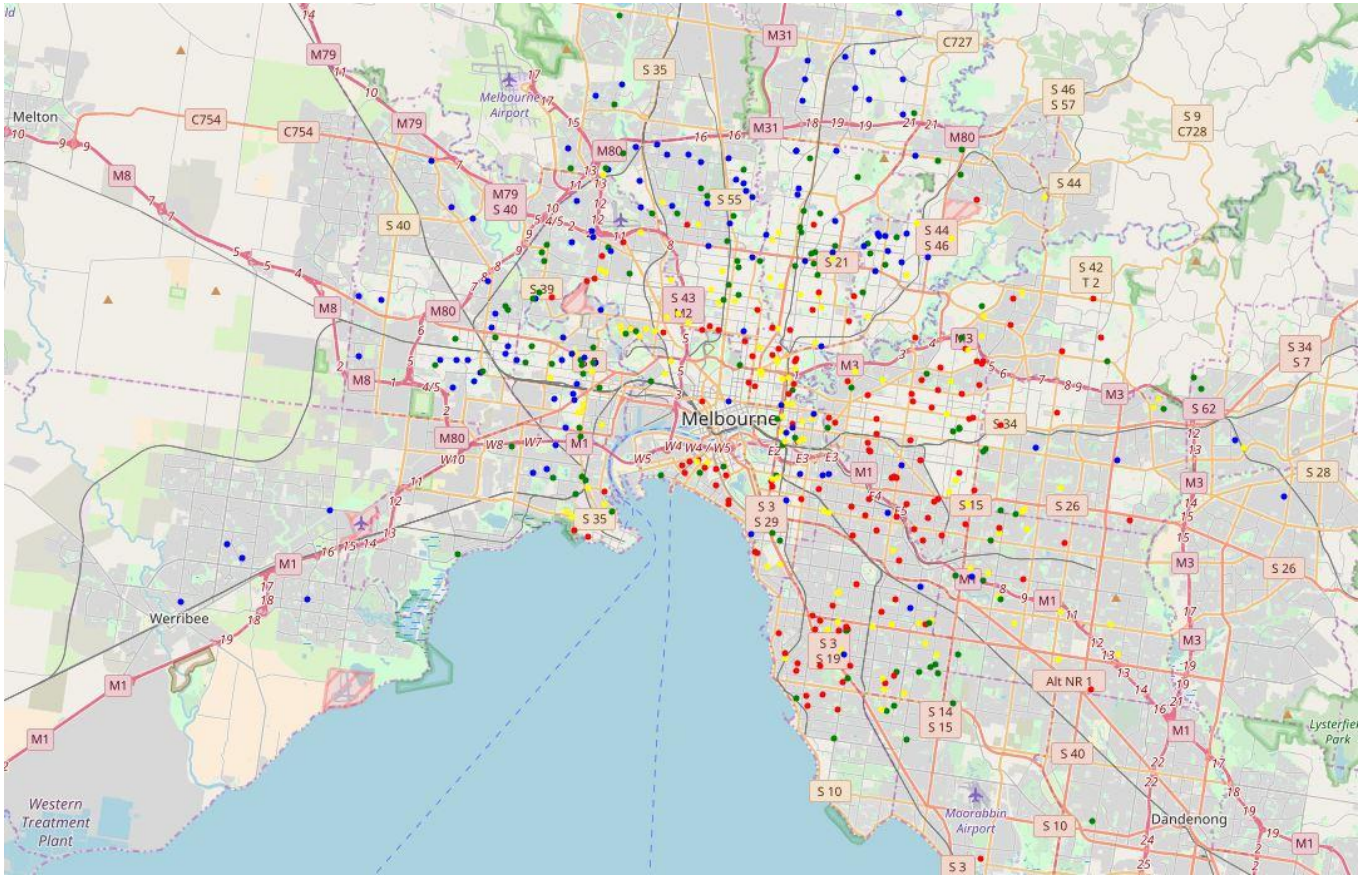
Variables	
Price	
Suburb	Distance
Address	Bedroom2
Rooms	Bathroom
Type	Car
Method	Landsize
SellerG	BuildingArea
Date	CouncilArea
YearBuilt	Lattitude
Regionname	Longtitude

- **유사한 변수 고려** : Suburb (Postcode / Regionname), Rooms (Bedroom2)
- **분석에서 제외된 변수** : BuildingArea(57.9%), YearBuilt(51.6%), Address, Date
- **분석에 포함된 변수 개수 :** 10,272개 (Train : 8,217개 / Test : 2,055개)

LG CNS

2. 데이터 전처리 및 EDA

Price를 기준으로 공간시각화를 통한 EDA를 진행함



Price 범위	색
1,400,000~	빨강
967,500~1,400,000	노랑
710,000 ~ 967,500	초록
0 ~ 710,000	파랑

Price의 범위는 사분위 값을 기준으로 나누었음

- Train데이터 중 200개의 sample을 위도, 경도 값을 이용하여 지도에 시각화함
- 북서쪽은 남동쪽에 비해 상대적으로 저렴한 가격대를 형성하고 있음을 확인함

LG CNS

3. 모델 적용 및 분석 - 1)RandomForest

Random Forest로 Price 예측을 하는 모델링을 수행함

분석에 사용한 모델
1) RandomForest
2) Support Vector Machine
3) Decision Tree + Regression

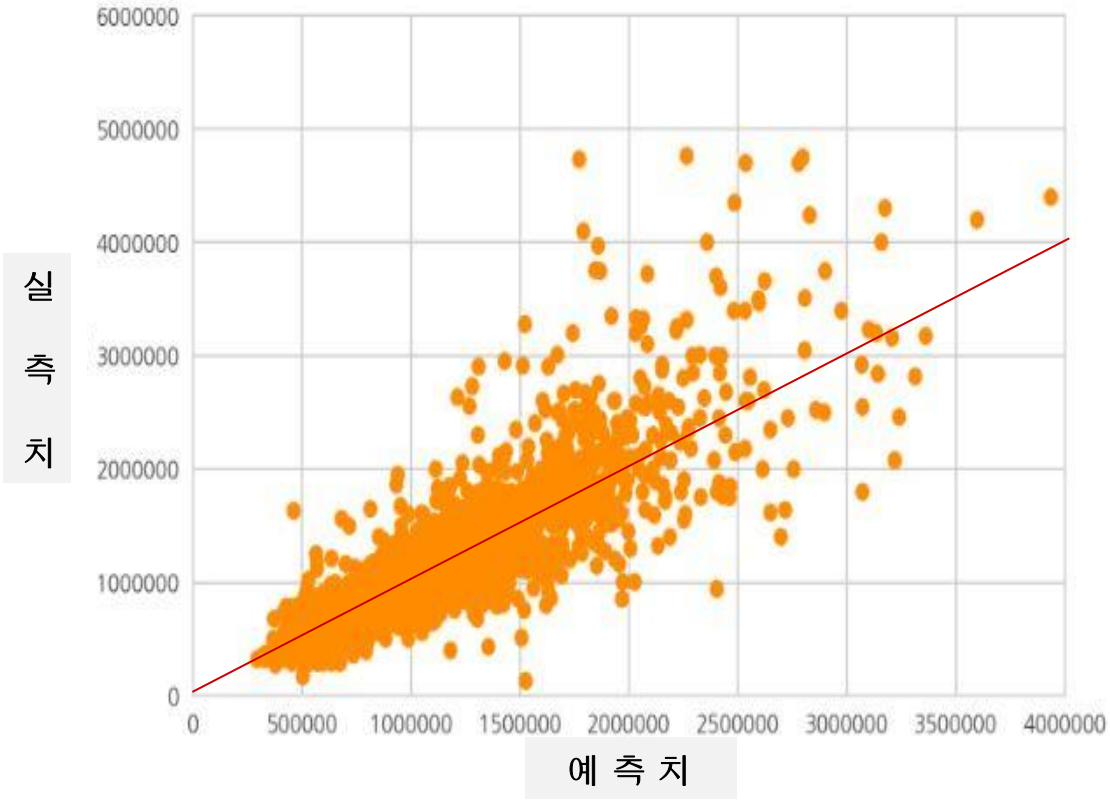
- Random Forest로 Price를 예측하는 모델링을 하였음
- Random Forest에 사용된 변수 : Suburb, Rooms, Type, Method, SellerG, Distance, Bathroom, Car, Landsize, CouncilArea, Regionname
- 범주형 변수는 encoding을 한 후에 Random Forest에 적용함
- Random Forest 모델의 RMSE와 Prediction Error값을 확인하였음

	RandomForest
RMSE	576,756
Prediction Error	0.269

LG CNS

3. 모델 적용 및 분석 - 2)SVM

SVM을 통하여 모델링하여 예측률을 비교함



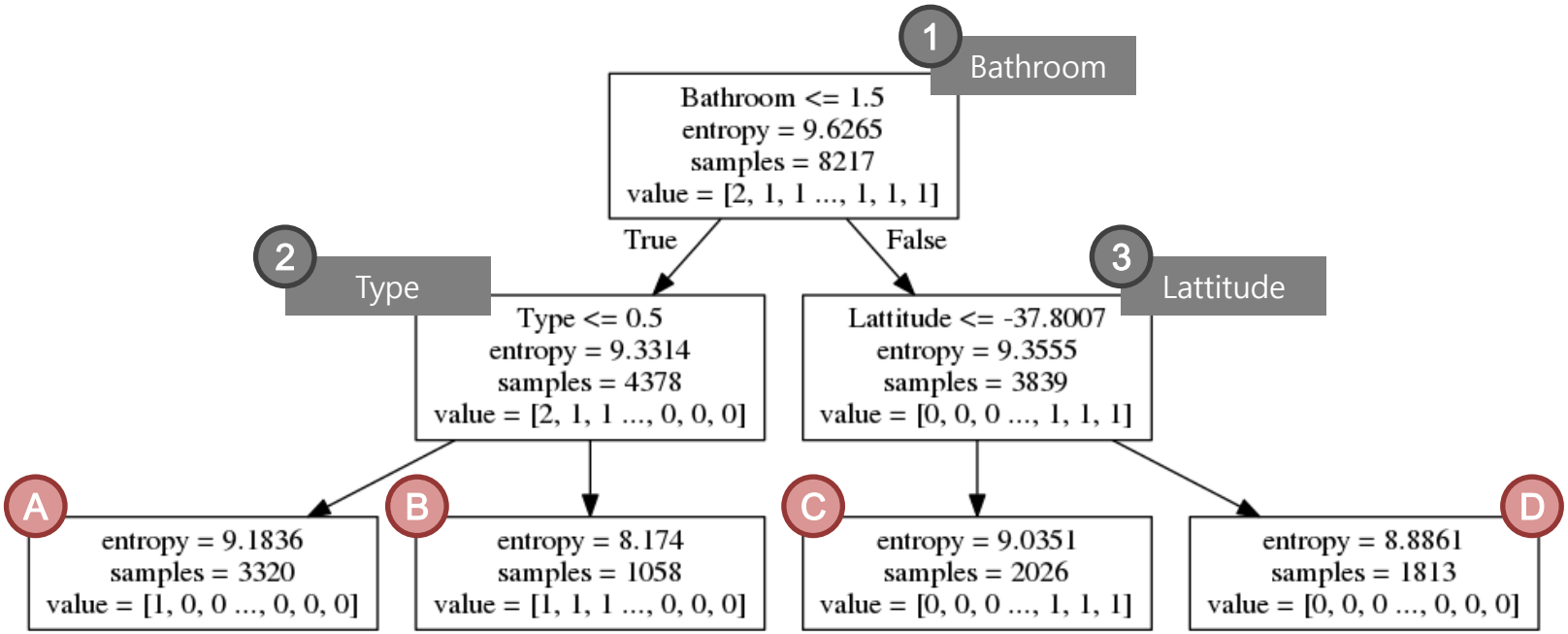
SVM	One hot encoding(Type)	
	전	후
RMSE	426,227	395,452
Prediction Error	0.201	0.187

- Support Vector Machine을 사용하여 Price 값을 예측함
- SVM에 독립변수로 Bathroom, Distance, Car, Landsize, Long, Latt, Type(one hot encoding)을 사용함
- 모든 범주형 변수에 one hot encoding을 적용하게 될 경우, 차원이 많이 늘어나기 때문에 다른 모델을 고민하게 되었음

LG CNS

3. 모델 적용 및 분석 - 3) Decision Tree / Regression

Decision Tree를 기반으로 데이터를 그룹화하여 그룹별 예측 회귀분석을 진행함

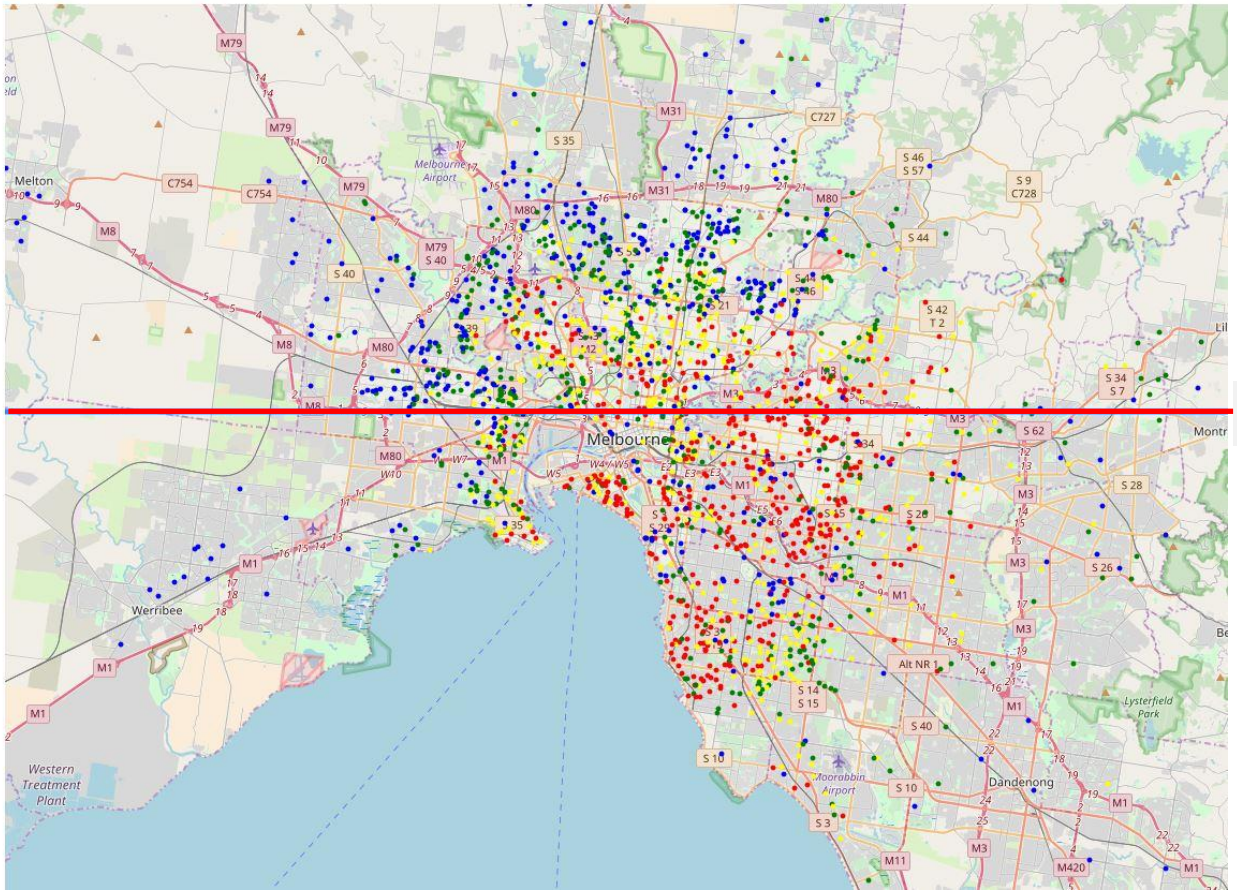


- Train data를 Decision Tree를 이용하여 4개의 그룹으로 나누었음
  (Sample의 개수 **group A** : 3320개, **group B** : 1058개, **group C**: 2026개, **group D**: 1813개)
- 변수 중에서 Bathroom, Type, Lattitude가 그룹을 나누는 것에 유의미한 변수로 결정됨
- 각 그룹별로 별도의 Regression 모델을 적용하여 분석을 하였음

LG CNS

3. 모델 적용 및 분석 - 3) Decision Tree / Regression

Decision Tree의 결과 변수 중 위도를 시각화를 통해 확인함



Lattitude : -37.8007

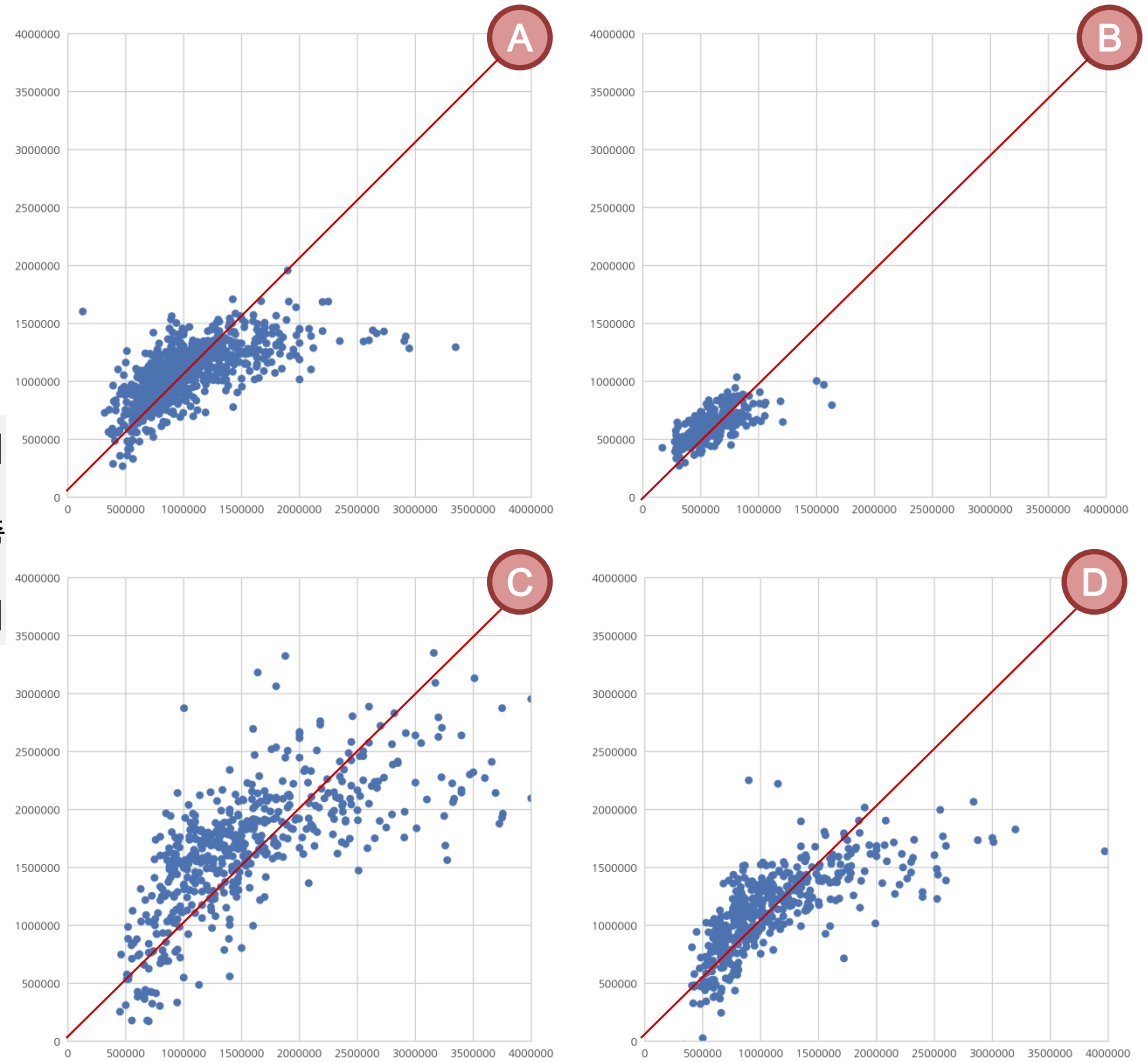
Price 범위	색
1,400,000~	빨강
967,500~1,400,000	노랑
710,000 ~ 967,500	초록
0 ~ 710,000	파랑

- Test데이터 2055개의 sample을 가격의 사분위 값을 기준으로 지도에 시각화함
- Decision Tree에서 분류된 기준대로, 위도 -37.8007을 중심으로 가격이 나뉘는 것을 볼 수 있음

LG CNS

3. 모델 적용 및 분석 - 3) Decision Tree / Regression

그룹별 regression모델의 Price 예측 결과를 비교 분석함



- Decision Tree를 통해 구분된 그룹별로 각각 Regression을 이용하여 분석을 함
- 상대적으로 Price가 낮은 변수들이 포함 된 group B의 경우 예측률이 가장 높았음

	RMSE	Pred.Error
A	416,445	0.255
B	156,822	0.202
C	625,174	0.308
D	403,731	0.277

LG CNS

4. Summary

모델 별 결과 비교

	Random Forest	SVM
RMSE	576,756	395,452
Prediction Error	0.269	0.187

	Decision Tree + Regression			
	Group A	Group B	Group C	Group D
RMSE	416,445	156,822	625,174	403,731
Prediction Error	0.255	0.202	0.308	0.277

모델 개선 가능성

- 결측치가 많은 변수 및 결측치를 포함한 데이터 처리 방법 (imputation 등)
- 유의미한 파생변수의 생성
- SVM에서 다양한 커널 사용
- 다양한 회귀분석 모델의 적용

LG CNS