Car Price Prediction

Kelompok 8

Anggota

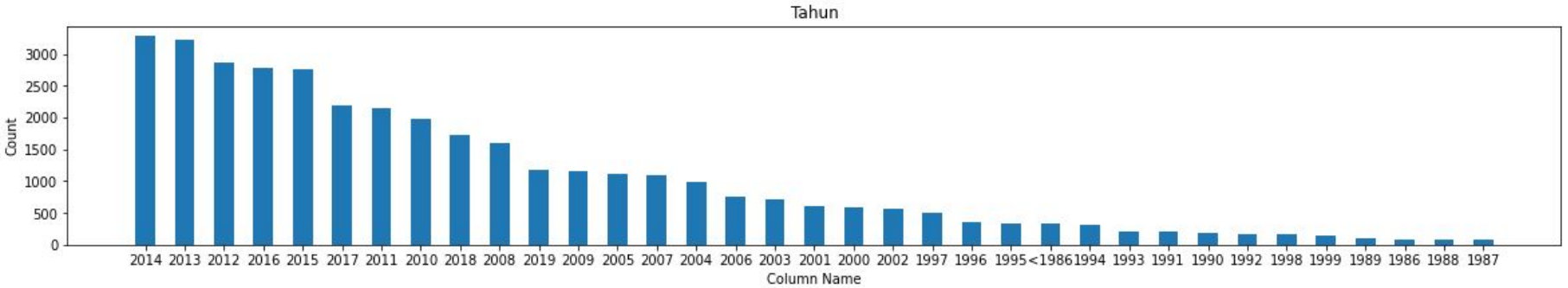
Nama	NIM	Task
Raihan Luthfi Haryawan	13517016	Pre-proses Data
Jesslyn Nathania	13517053	Model & Experiment
Adyaksa Wisanggeni	13517091	Model & Experiment
Edward Alexander Jaya	13517115	Data Analysis
Ridwan Faturrahman	13517150	Pre-proses Data, Eksplorasi Data, Model

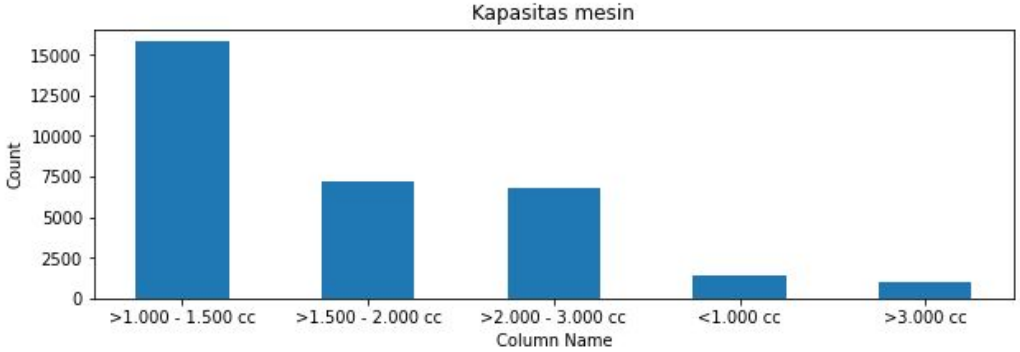
Background

Dalam melakukan regresi pada handson minggu ke-12, data yang digunakan adalah data harga mobil hasil *scrapping*. Data yang digunakan mengandung parameter berupa : `Tahun, Kapasitas mesin, Warna, Tipe bodi, Varian, STATE, Merek, Transmisi, Model, Fitur tambahan, Nama Bursa Mobil, Tipe bahan bakar, Tipe Penjual, CITY, COUNTRY, Jarak tempuh, phone, Sistem Penggerak, price,` dan `NEIGHBOURHOOD`.

Data Analysis

Berikut adalah beberapa analisis dari data yang didapat.

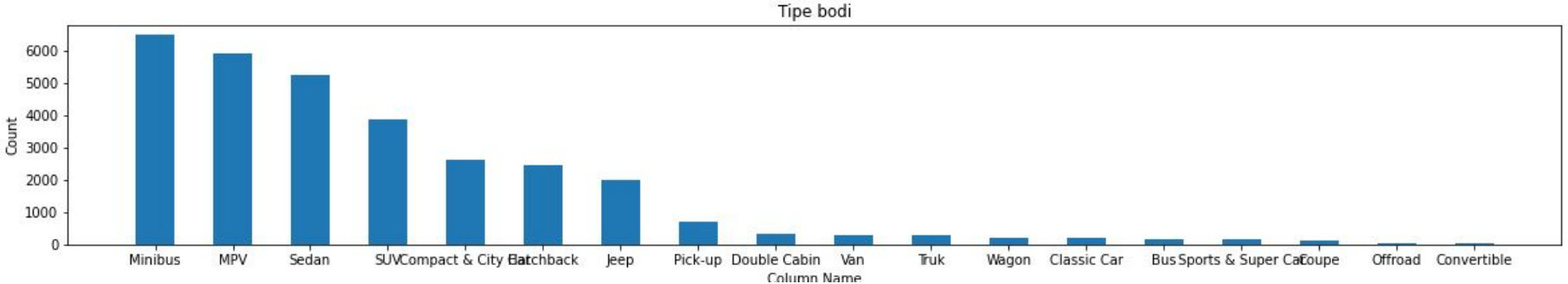

Tahun
Count
3000
2500
2000
1500
1000
500
0
2014 2013 2012 2016 2015 2017 2011 2010 2018 2008 2019 2009 2005 2007 2004 2006 2003 2001 2000 2002 1997 1996 1995 <1986 1994 1993 1991 1990 1992 1998 1999 1989 1986 1988 1987
Column Name


Kapasitas mesin
Count
15000
12500
10000
7500
5000
2500
0
>1.000 - 1.500 cc
>1.500 - 2.000 cc
>2.000 - 3.000 cc
<1.000 cc
>3.000 cc
Column Name

Tahun dan Kapasitas Mesin

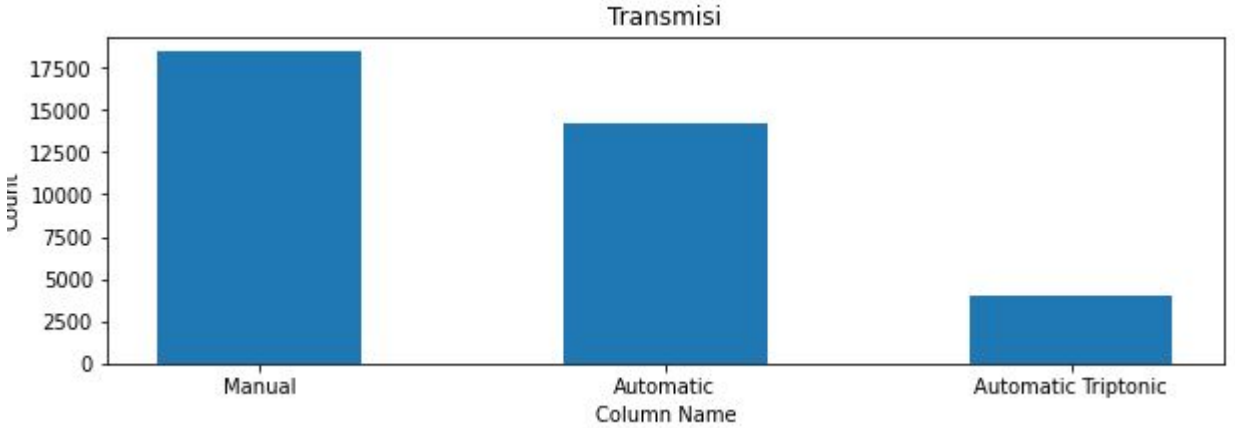
- Terdapat penurunan penjualan mobil seiring tahun (sedikit mobil yang dijual tahun 1990-an)
- Mobil yang dijual paling banyak memiliki kapasitas mesin di antara 1000 cc dan 1500 cc
- Sedikit penjualan mobil dengan kapasitas mesin di bawah 1000 cc dan di atas 3000 cc

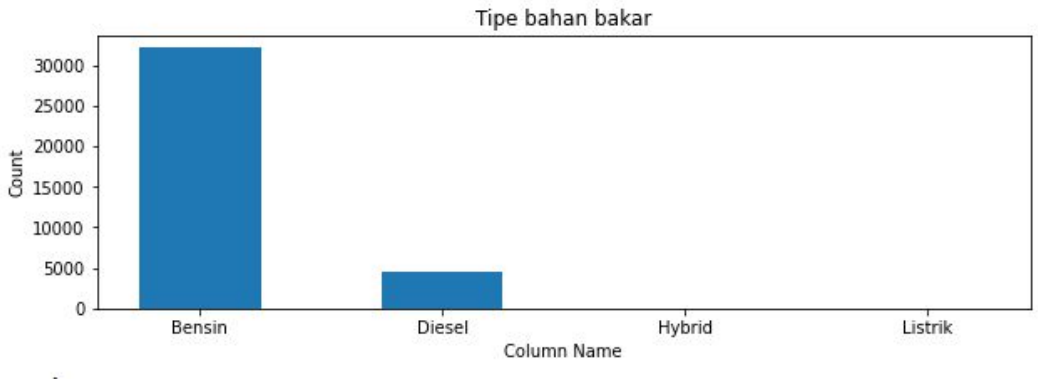

Warna
Count
10000
8000
6000
4000
2000
0
Hitam
Putih
Silver
Abu-abu
Merah
Biru
Hijau
Lainnya
Coklat
Marun
Kuning
Oranye
Emas
Ungu
Column Name

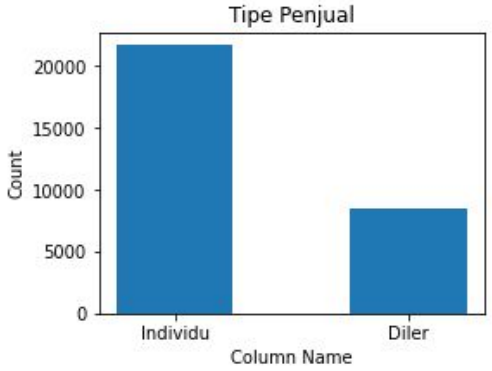

Tipe bodi
Count
6000
5000
4000
3000
2000
1000
0
Minibus
MPV
Sedan
SUV
Compact & City Car
Hatchback
Jeep
Pick-up
Double Cabin
Van
Truk
Wagon
Classic Car
Bus
Sports & Super Car
Coupe
Offroad
Convertible
Column Name

Warna dan Tipe Bodi

- Warna hitam, putih, dan silver mendominasi, namun sangat sedikit mobil dengan warna kuning, oranye, emas, dan ungu
- Tipe Bodi offroad dan convertible sangat sedikit (dapat diabaikan dari dataset)


Transmisi
17500
15000
12500
10000
7500
5000
2500
0
Count
Manual
Automatic
Automatic Triptonic
Column Name


Tipe bahan bakar
30000
25000
20000
15000
10000
5000
0
Count
Bensin
Diesel
Hybrid
Listrik
Column Name


Tipe Penjual
20000
15000
10000
5000
0
Count
Individu
Diler
Column Name

Transmisi, Tipe Bahan Bakar, Penjual

- Distribusi transmisi cukup baik
- Tipe bahan bakar hybrid dan listrik sangat sedikit (dapat diabaikan dari dataset)
- Distribusi tipe penjual cukup baik

Model

Pada handson ini, kami mencoba menggunakan beberapa model berikut.

1. Polynomial Linear Regression. Evaluation Metrics: R Squared (cocok untuk data
2. Polynomial Ridge Regression. Evaluation Metrics: R Squared
3. Lasso. Evaluation Metrics: R Squared
4. ElasticNet. Evaluation Metrics: R Squared
5. XGBoostRegressor. Evaluation Metrics: R Squared

Experiment

Eksperimen diterapkan pada enam skema dataset dengan tiap konfigurasi yang berbeda. Untuk menilai hasil dari eksperimen, digunakan splitting data train:test sebesar 80%:20% dan selanjutnya menghitung skor R2 dari data tes tersebut.

Skema Dataset dalam Eksperimen

```
df_schema_1 = Schema(df_schema,
                     [True,False,False,False,False,False,False] ,
                     [True,True,False,False,False,False,False,False,
                     False,False,False,False,False,False,False,False
                     ])

df_schema_2 = Schema(df_schema,
                     [True,True,True,False,False,False,False] ,
                     [True,True,True,False,False,False,False,False,
                     False,False,False,False,False,False,False,False
                     ])

df_schema_3 = Schema(df_schema,
                     [True,True,True,False,False,False,False] ,
                     [True,2,True,False,False,False,False,False,
                     False,False,False,False,False,False,False,False
                     ])
```

```
list_null_column = [
  'Kapasitas mesin',
  'Tipe bodi',
  'Varian',
  'Nama Bursa Mobil',
  'Tipe Penjual',
  'Sistem Penggerak',
  'NEIGHBOURHOOD'
]
```

True then drop row
False then replace with unknown value

```
df_schema_4 = Schema(df_schema,
                     [True,True,True,False,False,False,False] ,
                     [True,2,True,False,None,False,False,False,
                      False,False,False,False,False,None,False,None
                     ])

df_schema_5 = Schema(df_schema,
                     [True,True,True,False,False,False,False] ,
                     [True,2,True,False,None,False,None,False,False,
                      False,False,False,False,None,False,None
                     ])

df_schema_6 = Schema(df_schema,
                     [True,True,True,False,False,False,False] ,
                     [True,2,True,False,None,False,None,None,False,
                      False,False,False,False,None,False,None
                     ])
```

```
list_encoding_column = [
  'Kapasitas mesin', 'Tahun',
  'Warna', 'Tipe bodi', 'Varian',
  'Jarak tempuh',
  'Nama Bursa Mobil',
  'STATE', 'Merek', 'Transmisi',
  'Model', 'Tipe bahan bakar',
  'Tipe Penjual',  'CITY',
  'Sistem Penggerak',
  'NEIGHBOURHOOD'
]
```

True then one_hot_encoding
False then numerical_encoding
None then replace values with 0(zero)
2 then do nothing

Hasil experiment

	Poly Lin Deg-1	Poly Lin Deg-2	Poly Ridge Deg-1	Poly Ridge Deg-2	Lasso	ElasticNet	XGB
Schema-0	0.244630	-349.979351	0.244582	0.431524	0.244630	0.201392	0.751735
Schema-1	0.351045	-372780.170173	0.351091	0.488122	0.351047	0.307404	0.861882
Schema-2	0.338761	0.501371	0.338792	0.481954	0.338761	0.323156	0.860130
Schema-3	0.337218	0.475783	0.337251	0.474045	0.337218	0.321670	0.878129
Schema-4	0.333650	0.490146	0.333684	0.467849	0.333650	0.318372	0.894906
Schema-5	0.334079	0.493141	0.334115	0.468761	0.334079	0.319048	0.890747

Kesimpulan

- Secara general, XGBRegressor lebih baik dibandingkan regresi pada pangkat 1 dan 2
- Regresi dengan orde yang lebih tinggi belum tentu menghasilkan nilai yang lebih baik