Pacmann

Salary Prediction Based on Country and Race Using Regression Model

Statistics for Business

Jamaluddin Tuankotta

Outline

- Introduction
- Dataset
- Statistical Test
- Regression Model
- Conclusion and Recommendation
- References

Introduction

Introduction

Terdapat sebuah dataset yang menunjukkan gaji yang dimiliki seseorang berdasarkan usia, jenis kelamin, tingkat pendidikan, jabatan, ras, asal negara dan lama pengalaman kerja. Dari dataset tersebut, penulis ingin mengetahui pengaruh dari faktor-faktor tersebut terhadap gaji dan melakukan prediksi gaji seseorang.

- Menguji pengaruh jenis kelamin terhadap gaji dengan uji statistik.
- Memprediksi gaji dari lama pengalaman kerja seseorang dengan model regresi.
- Memprediksi gaji dari variabel prediktor usia, jenis kelamin, tingkat pendidikan, ras, asal negara dan lama pengalaman kerja dengan model regresi.

Dataset

Dataset

- Dataset yang digunakan diambil dari kaggle.com. Dataset berisikan 6704 baris data usia, jenis kelamin, tingkat pendidikan, jabatan, lama pengalaman kerja, ras, asal negara dan besar gaji.
- Sebelum diolah lebih lanjut, dilakukan persiapan dengan menghapus missing value dan duplicated data sehingga didapatkan 6698 baris data yang bisa digunakan.
- Data Job Title tidak akan digunakan dalam pemodelan regresi karena terlalu bervariasi.

Dataset

- Data jenis kelamin (Gender) akan diubah dari data kategorikal menjadi data numerik dengan Male = 0 dan Female = 1.
- Data tingkat pendidikan (Education Level) akan diubah dari data kategorikal menjadi data numerik dengan Bachelor's = 0, Master's = 1, PhD = 2, dan High School = 3.
- Data asal negara (Country) akan diubah dari data kategorikal menjadi data numerik dengan USA = 0, China = 1, dan Australia = 2, UK = 3, Canada = 4.
- Data ras (Race) akan diubah dari data kategorikal menjadi data numerik dengan White = 0, Asian = 1, dan Black = 2, Mixed = 3, Hispanic = 4.

Deskripsi Data

Deskripsi Data Numerik

	count	mean	std	min	25%	50%	75%	max
Age	6684.0	33.610563	7.595994	21.0	28.0	32.0	38.0	62.0
YearsOfExperience	6684.0	8.077723	6.029305	0.0	3.0	7.0	12.0	34.0
Salary	6684.0	115307.175194	52806.810881	350.0	70000.0	115000.0	160000.0	250000.0

Korelasi Antar Data Numerik

```
df_salary[["Age","YearsOfExperience", "Salary"]].corr()
```

	Age	YearsOfExperience	Salary
Age	1.000000	0.937609	0.728343
YearsOfExperience	0.937609	1.000000	0.810942
Salary	0.728343	0.810942	1.000000

Korelasi antara usia, lama pengalaman kerja, dan gaji memiliki hasil positif dan berkorelasi kuat.

Deskripsi Data

Deskripsi Data Kategorik

```
[ ] df_salary["Gender"].value_counts()

Male      3671
Female    3013
Name: Gender, dtype: int64
```

```
[ ] df_salary["Country"].value_counts()

USA          1356
China        1339
Australia    1335
UK           1332
Canada       1322
Name: Country, dtype: int64
```

```
[ ] df_salary["Race"].value_counts()

White       2742
Asian       2499
Black        787
Mixed        334
Hispanic     322
Name: Race, dtype: int64
```

Perbandingan Gaji Antar Variabel Kategorik

```
[87] df_salary.groupby("Gender")["Salary"].mean()

Gender
Female    107888.998672
Male      121395.697630
Name: Salary, dtype: float64
```

```
[88] df_salary.groupby("EducationLevel")["Salary"].mean()

EducationLevel
Bachelor's      95082.908640
High School     34415.612385
Master's       130078.384822
PhD            165651.457999
Name: Salary, dtype: float64
```

Deskripsi Data

Perbandingan Gaji Antar Variabel Kategorik

```
[90] df_salary.groupby("Country")["Salary"].mean()

Country
Australia    114925.465169
Canada       116455.090772
China        116282.589993
UK           115919.916667
USA          112998.758112
Name: Salary, dtype: float64
```

```
[91] df_salary.groupby("Race")["Salary"].mean()

Race
Asian       114876.360544
Black       115705.029225
Hispanic    110410.139752
Mixed       116330.859281
White       116035.997082
Name: Salary, dtype: float64
```

- Rata-rata gaji laki-laki lebih besar dari rata-rata gaji perempuan.
- Rata-rata gaji semakin besar seiring dengan level pendidikan yang lebih tinggi.
- Rata-rata gaji orang dari USA paling kecil dibandingkan negara lainnya.
- Rata-rata gaji orang dengan ras hispanik paling kecil dibandingkan ras lainnya.

Uji Statistik

Penulis ingin mengetahui pengaruh jenis kelamin terhadap besarnya gaji seseorang. Dalam dataset terdapat 2 jenis kelamin yaitu male (laki-laki (a)) dan female (perempuan (b)).
Penulis akan menguji apakah rata-rata gaji laki-laki lebih besar dari rata-rata gaji perempuan.

$$H_0: \mu_a = \mu_b$$

$$H_1: \mu_a > \mu_b$$

Karena standar deviasi populasi tidak diketahui, digunakan t-test. Sebelum menggunakan t-test, dilakukan uji variansi.

```
[40] # Gaji Laki-laki
     df_male = df_salary[df_salary["Gender"]=="Male"]["Salary"].values

     # Gaji Perempuan
     df_female = df_salary[df_salary["Gender"]=="Female"]["Salary"].values

     # Variansi
     np.var(df_male), np.var(df_female)

     (2713527473.797431, 2778856460.0537653)
```

Dari hasil tersebut dapat disimpulkan bahwa variansi tidak sama.

Uji Statistik

```
[41] from scipy import stats
     result = stats.ttest_ind(a = df_male,
                              b = df_female,
                              equal_var=False,
                              alternative = "greater")
```

```
[42] result.pvalue
```

```
8.848221350329085e-26
```

```
[43] result.statistic
```

```
10.476997992981117
```

```
# Menentukan aturan keputusan
if result.pvalue<significance_level:
    print("Tolak hipotesis nol.")
else:
    print("Gagal menolak hipotesis nol.")
```

```
Tolak hipotesis nol.
```

Terdapat cukup bukti bahwa rata-rata gaji laki-laki dan perempuan tidak sama.

Rata-rata gaji laki-laki lebih tinggi dan rata-rata gaji perempuan.

Pemodelan Regresi

Single Predictor

Dilakukan pemodelan regresi untuk memprediksi gaji seseorang dari lama pengalaman kerjanya.

```
[58] # Create OLS model object
     model = smf.ols("Salary ~ YearsOfExperience", df_salary)

     # Fit the model
     results_model_salary = model.fit()

     # Extract the results (Coefficient and Standard Error) to DataFrame
     results_salary = print_coef_std_err(results_model_salary)
     results_salary
```

	coef	std err
Intercept	57935.013495	631.916170
YearsOfExperience	7102.516670	62.693093

```
0s  results_model_salary.rsquared

    0.657626237954604
```

Dari hasil tersebut didapatkan persamaan regresi berikut dengan R-squared sebesar 0,65. Ini menunjukkan bahwa model dapat menjelaskan sekitar 65% variabilitas dari data.

Pemodelan Regresi

$$\text{Salary} = 57935 + 7102 \times \text{Years of Experience}$$

- Jika membandingkan dua orang yang memiliki 1 tahun perbedaan pada lama pengalaman kerja, diperkirakan orang yang memiliki pengalaman kerja lebih lama memiliki gaji yang lebih besar dengan selisih 7102.

- Untuk seseorang yang memiliki lama pengalaman kerja 0 tahun, perkiraan rata-rata gaji yang didapatkan adalah sebesar 57935.

Pemodelan Regresi

Single Predictor with Log Transformation

Dilakukan pemodelan regresi untuk memprediksi gaji seseorang dari lama pengalaman kerjanya namun dilakukan transformasi logaritmik pada variabel prediktor.

```
[65] # Create OLS model object
     model = smf.ols("Salary ~ logYOE", df_salary)

     # Fit the model
     results_logtransform = model.fit()

     # Extract the results (Coefficient and Standard Error) to DataFrame
     results_salary_log = print_coef_std_err(results_logtransform)
     results_logtransform.rsquared

     0.7116310655293663
```

Didapatkan hasil R-squared sebesar 0,71. Hasil ini lebih tinggi dari hasil R-squared 0,65 pada pemodelan tanpa transformasi logaritmik.

Sehingga untuk pemodelan regresi dengan satu variabel prediktor, digunakan model regresi dengan transformasi.

Pemodelan Regresi

Multiple Predictors with One Interaction

Dalam pemodelan ini, digunakan semua variabel prediktor yaitu usia, jenis kelamin, tingkat pendidikan, dan lama pengalaman kerja. Ditambahkan juga satu interaksi antar variabel prediktor yaitu usia dan lama pengalaman kerja. Untuk variabel tingkat pendidikan (Education Level) diperlakukan sebagai variabel kategorikal.

Didapatkan R-squared rata-rata sebesar 0,75 yang berarti model ini baik dan dapat menjelaskan 75% variansi gaji.

Evaluasi model dengan K-Fold cross validation

	test_rsquared	folds
0	0.759227	Folds 1
1	0.739663	Folds 2
2	0.746412	Folds 3
3	0.765434	Folds 4
4	0.741552	Folds 5

```
[69] scores_ols_all_pred["test_rsquared"].mean()

0.7504575523323711
```

Fitting Model

	coef	std err
Intercept	53250.822023	3737.948616
C(EducationLevel)[T.1]	5847.019743	864.247971
C(EducationLevel)[T.2]	20059.805197	1049.646099
C(EducationLevel)[T.3]	-25058.519605	1425.260561
Age	-235.204851	137.459072
Gender	-5393.177226	665.088171
Race	-254.486388	302.823858
Country	300.855820	230.364034
YearsOfExperience	16172.099090	311.394978
Age:YearsOfExperience	-199.395766	6.637300

Didapatkan hasil koefisien persamaan regresi sebagai di samping. Untuk mendapatkan intrepetasi yang lebih bermakna dan mudah, maka dilakukanlah centering by mean terhadap variabel usia (age).

Centering Variabel Usia (Age)

```
[ ]  mean_age = df_salary["Age"].mean()
     mean_age = np.round(mean_age,0)
     mean_age

     34.0
```

Digunakan rata-rata usia pada dataset (34 tahun) sebagai acuan. Sehingga data usia akan dihitung dari jaraknya terhadap usia 34 tahun.

Didapatkan R-squared rata-rata sebesar 0,75 yang berarti model ini baik dan dapat menjelaskan 75% variansi gaji.

Evaluasi model dengan K-Fold cross validation

	test_rsquared	folds
0	0.759313	Folds 1
1	0.759307	Folds 2
2	0.749453	Folds 3
3	0.739293	Folds 4
4	0.744070	Folds 5

```
[ ]  scores_ols_all_pred["test_rsquared"].mean()

     0.7502874434492411
```

Intrepretasi

Interpretasi tingkat pendidikan
Jika membandingkan dua orang yang memiliki usia, jenis kelamin, lama pengalaman kerja yang sama, gaji seseorang dengan tingkat pendidikan Bachelor's diperkirakan lebih tinggi 5847 dollar daripada gaji seseorang dengan tingkat pendidikan High School.

Interpretasi usia
Jika membandingkan dua orang yang memiliki jenis kelamin dan tingkat pendidikan yang sama, serta pengalaman kerja 0 tahun, seseorang yang usianya 1 tahun lebih tua dari 34 tahun diperkirakan memiliki gaji lebih tinggi 235 dollar daripada seseorang berusia 34 tahun.

Interpretasi jenis kelamin
Jika membandingkan dua orang yang memiliki usia, lama pengalaman kerja, dan tingkat pendidikan yang sama, perempuan diperkiraan memiliki gaji lebih sedikit 5393 dollar dibandingkan laki-laki.

Intrepretasi

Interpretasi asal negara
Jika membandingkan dua orang yang memiliki usia, lama pengalaman kerja, dan tingkat pendidikan yang sama, USA diperkiraan memiliki gaji lebih sedikit 301 dollar dibandingkan asal negara Australia.

Intrepetasi ras
Jika membandingkan dua orang yang memiliki usia, lama pengalaman kerja, dan tingkat pendidikan yang sama, hispanik diperkiraan memiliki gaji lebih sedikit 254 dollar dibandingkan ras asian.

Interpretasi lama pengalaman kerja
Jika membandingkan dua orang berusia 34 tahun yang memiliki jenis kelamin dan tingkat pendidikan yang sama, seseorang dengan lama pengalaman kerja lebih lama 1 tahun diperkirakan memiliki gaji lebih tinggi 199 dollar.

Conclusion and Recommendations

Conclusion

- Dapat disimpulkan bahwa usia, jenis kelamin, lama pengalaman kerja, dan tingkat pendidikan, asal negara dan ras berpengaruh terhadap besaran gaji seseorang sehingga bisa digunakan untuk memprediksi besaran gaji tersebut.

- Model regresi yang dibangun dengan single predictor yaitu lama pengalaman kerja menghasilkan performa yang cukup bagus dengan R-squared 0,65.

- Transformasi logaritmik pada model ini menghasilkan model yang lebih baik karena memiliki skor R-squared lebih tinggi yaitu 0,71.

- Model regresi yang dibangun dengan semua predictor disertai interaksi antara usia dan lama pengalaman kerja, menghasilkan performa yang lebih baik dengan R-squared 0,75. Model tersebut juga menghasilkan interpretasi yang baik dengan dilakukannya centering pada variabel usia (age).

Recommendation

- Untuk pengembangan selanjutnya dapat dilakukan percobaan untuk berbagai variasi jumlah predictor yang digunakan. Dapat juga dilakukan pengelompokan data gaji berdasarkan job title-nya sehingga didapatkan model regresi yang akurat untuk masing-masing jenis pekerjaan.

Reference

- Statistics for Business : Decision Making and Analysis—Robert Stine and Dean Foster
- Regression and Other Stories.—Andrew Gelman, Jennifer Hill, and Aki Vehtari
- The Effect: An Introduction to Research Design and Causality. Chapter 13 Huntington-Klein, N. 2021
- Stats: Salary Prediction Using Regression Model https://kristalinaks.medium.com/stats-salary-prediction-using-regression-model-f5e53c254ac6

Thank You