DS salary Prediction

2023-10-29

```
ds_dataset <- read.csv("salaries.csv")
head(ds_dataset)
```

```
##   work_year experience_level employment_type                 job_title salary
## 1      2023               SE              FT            Data Scientist 199000
## 2      2023               SE              FT            Data Scientist 196760
## 3      2023               SE              FT Machine Learning Engineer  90000
## 4      2023               SE              FT Machine Learning Engineer  70000
## 5      2023               SE              FT               ML Engineer 324000
## 6      2023               SE              FT               ML Engineer 159000
##   salary_currency salary_in_usd employee_residence remote_ratio
## 1             USD        199000                 US            0
## 2             USD        196760                 US            0
## 3             USD         90000                 CO            0
## 4             USD         70000                 CO            0
## 5             USD        324000                 US            0
## 6             USD        159000                 US            0
##   company_location company_size
## 1               US            M
## 2               US            M
## 3               CO            M
## 4               CO            M
## 5               US            M
## 6               US            M
```

```
nrow(ds_dataset)
```

```
## [1] 8113
```

```
ds_df <- ds_dataset %>%
         na.omit %>%
         select(-salary, -salary_currency)

categorical_vars <- c('work_year', 'experience_level', 'employment_type',
                      'job_title', 'remote_ratio', 'company_location',
                      'company_size')
north_america <- c('CA', 'US', 'PR')
south_america <- c('AR', 'BR', 'CL', 'CO', 'EC', 'HN', 'MX', 'PE')
europe <- c('AD', 'AE', 'AM', 'AS', 'AT', 'BA', 'BE', 'BS', 'CF', 'CH', 'CZ',
            'DE', 'DK', 'DZ', 'EE', 'EG', 'ES', 'FI', 'FR', 'GB', 'GH', 'GR',
            'HR', 'HU', 'IE', 'IL', 'IT', 'LT', 'LU', 'LV', 'MD', 'MT', 'MU',
            'NL', 'NO', 'PL', 'PT', 'RO', 'RU', 'SE', 'SI', 'UA')
asia <- c('CN', 'HK', 'ID', 'IN', 'IQ', 'IR', 'JP', 'KR', 'MY', 'PH', 'PK',
          'SA', 'SG', 'TH', 'TR', 'QA')
africa <- c('KE', 'NG', 'ZA')
oceania <- c('AU', 'NZ')

ds_df <- ds_df %>%
        mutate(
          job_title = case_when(
            grepl("Machine Learn", job_title) ~ "ML",
            grepl("Data Scie|Applied|Model", job_title) ~ "Data Scientists",
            grepl("Data Anal", job_title) ~ "Data Analytics",
            grepl("Data Visual|Power", job_title) ~ "Data Visualization",
            grepl("Architect", job_title) ~ "Data Architect",
            grepl("Decision|Strategy|Insight|Consultant", job_title) ~
              "Data consultant",
            grepl("AI", job_title) ~ "AI",
            grepl("Cloud", job_title) ~ "Cloud",
            grepl("Engin|ETL", job_title) ~ "Engineer",
            grepl("BI|Business", job_title) ~ "Business Intelligence",
            grepl("Research", job_title) ~ "Research",
            grepl("Specia", job_title) ~ "Specialist",
            grepl("Manage", job_title) ~ "Manager",
            TRUE ~ job_title),
          company_location = case_when(
            company_location %in% north_america ~ 'NorthAmerica',
            company_location %in% south_america ~ 'SouthAmerica',
            company_location %in% europe ~ 'Europe',
            company_location %in% asia ~ 'Aisa',
            company_location %in% africa ~ 'Africa',
            company_location %in% oceania ~ 'Oceania',
            TRUE ~ company_location),
          across(categorical_vars, as.factor)) %>%
          select(-employee_residence)
head(ds_df)
```

```
##   work_year experience_level employment_type       job_title salary_in_usd
## 1      2023               SE              FT Data Scientists        199000
## 2      2023               SE              FT Data Scientists        196760
## 3      2023               SE              FT              ML         90000
## 4      2023               SE              FT              ML         70000
## 5      2023               SE              FT        Engineer        324000
## 6      2023               SE              FT        Engineer        159000
##   remote_ratio company_location company_size
## 1            0     NorthAmerica            M
## 2            0     NorthAmerica            M
## 3            0     SouthAmerica            M
## 4            0     SouthAmerica            M
## 5            0     NorthAmerica            M
## 6            0     NorthAmerica            M
```

```
ds_df %>%
    group_by(work_year) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 4 x 2
##   work_year     avg
##   <fct>       <dbl>
## 1 2020      102251.
## 2 2021       99922.
## 3 2022      134508.
## 4 2023      155579.
```

```
ds_df %>%
    group_by(experience_level) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 4 x 2
##   experience_level     avg
##   <fct>              <dbl>
## 1 EN                85940.
## 2 EX               189670.
## 3 MI               114514.
## 4 SE               161643.
```

```
ds_df %>%
    group_by(employment_type) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 4 x 2
##   employment_type     avg
##   <fct>             <dbl>
## 1 CT              120838.
## 2 FL               54734.
## 3 FT              149654.
## 4 PT               52053.
```

```
ds_df %>%
    group_by(job_title) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 20 x 2
##    job_title                         avg
##    <fct>                           <dbl>
##  1 AI                            131609.
##  2 Autonomous Vehicle Technician  82778.
##  3 Business Intelligence         120063.
##  4 Cloud                         144608.
##  5 Data Analytics                109492.
##  6 Data Architect                167330.
##  7 Data consultant               144442.
##  8 Data Developer                103738.
##  9 Data Lead                     176500
## 10 Data Operations Analyst        92899
## 11 Data Quality Analyst           93324.
## 12 Data Scientists               160743.
## 13 Data Strategist                95938.
## 14 Data Visualization            116889.
## 15 Engineer                      150888.
## 16 Head of Data                  209119.
## 17 Manager                       109716.
## 18 ML                            177970.
## 19 Research                      171883.
## 20 Specialist                     94151.
```

```
ds_df %>%
    group_by(remote_ratio) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 3 x 2
##   remote_ratio     avg
##   <fct>          <dbl>
## 1 0            155719.
## 2 50            82441.
## 3 100          144149.
```

```
ds_df %>%
    group_by(company_location) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 6 x 2
##   company_location     avg
##   <fct>              <dbl>
## 1 Africa            62771.
## 2 Aisa              50500.
## 3 Europe            88336.
## 4 NorthAmerica     158234.
## 5 Oceania          132700.
## 6 SouthAmerica      70982.
```

```
ds_df %>%
    group_by(company_size) %>%
    summarise(avg = mean(salary_in_usd))
```

```
## # A tibble: 3 x 2
##   company_size     avg
##   <fct>          <dbl>
## 1 L            133531.
## 2 M            152250.
## 3 S             88557.
```

```
full <- lm(salary_in_usd ~., ds_df)
summary(full)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ ., data = ds_df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -144754  -35235   -5962   27685  391031 
## 
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                             92523.5    21486.4   4.306 1.68e-05 ***
## work_year2021                          -9149.2     7005.0  -1.306 0.191559    
## work_year2022                          -8918.2     6380.0  -1.398 0.162203    
## work_year2023                           3119.2     6350.9   0.491 0.623343    
## experience_levelEX                      78811.0     4349.4  18.120  < 2e-16 ***
## experience_levelMI                      23704.9     2893.9   8.191 2.98e-16 ***
## experience_levelSE                      49955.2     2746.2  18.191  < 2e-16 ***
## employment_typeFL                      -53860.5    20239.5  -2.661 0.007803 ** 
## employment_typeFT                       -7304.7    12421.2  -0.588 0.556491    
## employment_typePT                       -3975.1    19020.0  -0.209 0.834456    
## job_titleAutonomous Vehicle Technician  17988.4    38343.9   0.469 0.638987    
## job_titleBusiness Intelligence         -54148.5     7920.4  -6.837 8.70e-12 ***
## job_titleCloud                          -3137.5    18401.7  -0.171 0.864620    
## job_titleData Analytics                -56303.5     6429.0  -8.758  < 2e-16 ***
## job_titleData Architect                -10739.2     7307.6  -1.470 0.141710    
## job_titleData consultant               -26477.7     9609.2  -2.755 0.005874 ** 
## job_titleData Developer                -65381.0    22106.8  -2.958 0.003110 ** 
## job_titleData Lead                     -13231.9    17586.2  -0.752 0.451833    
## job_titleData Operations Analyst       -73977.4    17570.4  -4.210 2.58e-05 ***
## job_titleData Quality Analyst          -90435.0    14429.8  -6.267 3.86e-10 ***
## job_titleData Scientists               -11246.2     6372.5  -1.765 0.077635 .  
## job_titleData Strategist               -87127.2    14425.1  -6.040 1.61e-09 ***
## job_titleData Visualization            -65544.2    16860.2  -3.888 0.000102 ***
## job_titleEngineer                      -21203.4     6361.8  -3.333 0.000863 ***
## job_titleHead of Data                   22322.1    11873.1   1.880 0.060136 .  
## job_titleManager                       -65782.5     7756.0  -8.481  < 2e-16 ***
## job_titleML                              5654.3     6481.3   0.872 0.383018    
## job_titleResearch                        1052.3     7080.5   0.149 0.881863    
## job_titleSpecialist                    -69819.1     9552.2  -7.309 2.94e-13 ***
## remote_ratio50                         -14075.8     4180.3  -3.367 0.000763 ***
## remote_ratio100                         -5552.9     1240.2  -4.478 7.66e-06 ***
## company_locationAisa                   -21252.9    16732.8  -1.270 0.204073    
## company_locationEurope                  -2935.7    15925.6  -0.184 0.853755    
## company_locationNorthAmerica            52858.5    15873.2   3.330 0.000872 ***
## company_locationOceania                 56807.8    19028.4   2.985 0.002840 ** 
## company_locationSouthAmerica           -23374.3    17391.8  -1.344 0.178990    
## company_sizeM                            -830.2     2275.2  -0.365 0.715216    
## company_sizeS                          -15634.4     4641.0  -3.369 0.000759 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 51860 on 8075 degrees of freedom
## Multiple R-squared:  0.3466, Adjusted R-squared:  0.3436 
## F-statistic: 115.8 on 37 and 8075 DF,  p-value: < 2.2e-16
```