Assignment 04

Yibei Yu

October 7, 2025

Github Link: https://github.com/met-ad-688/assignment-04-yibei23-jpg

“This homework was done with the help of AI for code debugging.”

# 1. Setup and Data Loading

[Stage 115:> (0 + 1) / 1]

+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
| ID|LAST\_UPDATED\_DATE|LAST\_UPDATED\_TIMESTAMP|DUPLICATES| POSTED| EXPIRED|DURATION| SOURCE\_TYPES| SOURCES| URL|ACTIVE\_URLS|ACTIVE\_SOURCES\_INFO| TITLE\_RAW| BODY|MODELED\_EXPIRED|MODELED\_DURATION| COMPANY| COMPANY\_NAME|COMPANY\_RAW|COMPANY\_IS\_STAFFING|EDUCATION\_LEVELS|EDUCATION\_LEVELS\_NAME|MIN\_EDULEVELS| MIN\_EDULEVELS\_NAME|MAX\_EDULEVELS|MAX\_EDULEVELS\_NAME|EMPLOYMENT\_TYPE|EMPLOYMENT\_TYPE\_NAME|MIN\_YEARS\_EXPERIENCE|MAX\_YEARS\_EXPERIENCE|IS\_INTERNSHIP|SALARY|REMOTE\_TYPE|REMOTE\_TYPE\_NAME|ORIGINAL\_PAY\_PERIOD|SALARY\_TO|SALARY\_FROM| LOCATION| CITY| CITY\_NAME|COUNTY| COUNTY\_NAME| MSA| MSA\_NAME|STATE|STATE\_NAME|COUNTY\_OUTGOING|COUNTY\_NAME\_OUTGOING|COUNTY\_INCOMING|COUNTY\_NAME\_INCOMING|MSA\_OUTGOING| MSA\_NAME\_OUTGOING|MSA\_INCOMING| MSA\_NAME\_INCOMING|NAICS2| NAICS2\_NAME|NAICS3| NAICS3\_NAME|NAICS4| NAICS4\_NAME|NAICS5| NAICS5\_NAME|NAICS6| NAICS6\_NAME| TITLE| TITLE\_NAME| TITLE\_CLEAN| SKILLS| SKILLS\_NAME| SPECIALIZED\_SKILLS|SPECIALIZED\_SKILLS\_NAME| CERTIFICATIONS| CERTIFICATIONS\_NAME| COMMON\_SKILLS| COMMON\_SKILLS\_NAME| SOFTWARE\_SKILLS|SOFTWARE\_SKILLS\_NAME| ONET| ONET\_NAME| ONET\_2019| ONET\_2019\_NAME| CIP6| CIP6\_NAME| CIP4| CIP4\_NAME| CIP2| CIP2\_NAME|SOC\_2021\_2| SOC\_2021\_2\_NAME|SOC\_2021\_3| SOC\_2021\_3\_NAME|SOC\_2021\_4|SOC\_2021\_4\_NAME|SOC\_2021\_5|SOC\_2021\_5\_NAME|LOT\_CAREER\_AREA|LOT\_CAREER\_AREA\_NAME|LOT\_OCCUPATION| LOT\_OCCUPATION\_NAME|LOT\_SPECIALIZED\_OCCUPATION|LOT\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_OCCUPATION\_GROUP|LOT\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_SPECIALIZED\_OCCUPATION|LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION|LOT\_V6\_OCCUPATION\_NAME|LOT\_V6\_OCCUPATION\_GROUP|LOT\_V6\_OCCUPATION\_GROUP\_NAME|LOT\_V6\_CAREER\_AREA|LOT\_V6\_CAREER\_AREA\_NAME| SOC\_2| SOC\_2\_NAME| SOC\_3| SOC\_3\_NAME| SOC\_4| SOC\_4\_NAME| SOC\_5| SOC\_5\_NAME|LIGHTCAST\_SECTORS|LIGHTCAST\_SECTORS\_NAME|NAICS\_2022\_2| NAICS\_2022\_2\_NAME|NAICS\_2022\_3| NAICS\_2022\_3\_NAME|NAICS\_2022\_4| NAICS\_2022\_4\_NAME|NAICS\_2022\_5| NAICS\_2022\_5\_NAME|NAICS\_2022\_6| NAICS\_2022\_6\_NAME|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
|1f57d95acf4dc67ed...| 9/6/2024| 2024-09-06 20:32:...| 0|6/2/2024| 6/8/2024| 6| [\n "Company"\n]|[\n "brassring.c...|[\n "https://sjo...| []| NULL|Enterprise Analys...|31-May-2024\n\nEn...| 6/8/2024| 6| 894731| Murphy USA| Murphy USA| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 2| 2| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.20...|RWwgRG9yYWRvLCBBUg==|El Dorado, AR| 5139| Union, AR|20980| El Dorado, AR| 5| Arkansas| 5139| Union, AR| 5139| Union, AR| 20980| El Dorado, AR| 20980| El Dorado, AR| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...|441330|Automotive Parts ...|ET29C073C03D1F86B4|Enterprise Analysts|enterprise analys...|[\n "KS126DB6T06...|[\n "Merchandisi...|[\n "KS126DB6T06...| [\n "Merchandisi...| []| []|[\n "KS126706DPF...|[\n "Mathematics...|[\n "KS440W865GC...|[\n "SQL (Progra...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...|[\n "45.0601",\n...|[\n "Economics, ...|[\n "45.06",\n ...|[\n "Economics",...|[\n "45",\n "27...|[\n "Social Scie...| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101011| General ERP Analy...| 2310| Business Intellig...| 23101011| General ERP Analy...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 7\n]| [\n "Artificial ...| 44| Retail Trade| 441|Motor Vehicle and...| 4413|Automotive Parts,...| 44133|Automotive Parts ...| 441330|Automotive Parts ...|  
|0cb072af26757b6c4...| 8/2/2024| 2024-08-02 17:08:...| 0|6/2/2024| 8/1/2024| NULL| [\n "Job Board"\n]| [\n "maine.gov"\n]|[\n "https://job...| []| NULL|Oracle Consultant...|Oracle Consultant...| 8/1/2024| NULL| 133098|Smx Corporation L...| SMX| true| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| 3| false| NULL| 1| Remote| NULL| NULL| NULL|{\n "lat": 44.31...| QXVndXN0YSwgTUU=| Augusta, ME| 23011| Kennebec, ME|12300|Augusta-Watervill...| 23| Maine| 23011| Kennebec, ME| 23011| Kennebec, ME| 12300|Augusta-Watervill...| 12300|Augusta-Watervill...| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...|561320|Temporary Help Se...|ET21DDA63780A7DC09| Oracle Consultants|oracle consultant...|[\n "KS122626T55...|[\n "Procurement...|[\n "KS122626T55...| [\n "Procurement...| []| []| []| []|[\n "BGSBF3F508F...|[\n "Oracle Busi...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 56|Administrative an...| 561|Administrative an...| 5613| Employment Services| 56132|Temporary Help Se...| 561320|Temporary Help Se...|  
|85318b12b3331fa49...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024| 7/7/2024| 35| [\n "Job Board"\n]|[\n "dejobs.org"\n]|[\n "https://dej...| []| NULL| Data Analyst|Taking care of pe...| 6/10/2024| 8|39063746| Sedgwick| Sedgwick| false| [\n 2\n]| [\n "Bachelor's ...| 2| Bachelor's degree| NULL| NULL| 1|Full-time (> 32 h...| 5| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 32.77...| RGFsbGFzLCBUWA==| Dallas, TX| 48113| Dallas, TX|19100|Dallas-Fort Worth...| 48| Texas| 48113| Dallas, TX| 48113| Dallas, TX| 19100|Dallas-Fort Worth...| 19100|Dallas-Fort Worth...| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...|524291| Claims Adjusting|ET3037E0C947A02404| Data Analysts| data analyst|[\n "KS1218W78FG...|[\n "Management"...|[\n "ESF3939CE1F...| [\n "Exception R...|[\n "KS683TN76T7...|[\n "Security Cl...|[\n "KS1218W78FG...|[\n "Management"...|[\n "KS126HY6YLT...|[\n "Microsoft O...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 52|Finance and Insur...| 524|Insurance Carrier...| 5242|Agencies, Brokera...| 52429|Other Insurance R...| 524291| Claims Adjusting|  
|1b5c3941e54a1889e...| 9/6/2024| 2024-09-06 20:32:...| 1|6/2/2024|7/20/2024| 48| [\n "Job Board"\n]|[\n "disabledper...|[\n "https://www...| []| NULL|Sr. Lead Data Mgm...|About this role:\...| 6/12/2024| 10|37615159| Wells Fargo|Wells Fargo| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 1|Full-time (> 32 h...| 3| NULL| false| NULL| 0| [None]| NULL| NULL| NULL|{\n "lat": 33.44...| UGhvZW5peCwgQVo=| Phoenix, AZ| 4013| Maricopa, AZ|38060|Phoenix-Mesa-Chan...| 4| Arizona| 4013| Maricopa, AZ| 4013| Maricopa, AZ| 38060|Phoenix-Mesa-Chan...| 38060|Phoenix-Mesa-Chan...| 52|Finance and Insur...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking|522110| Commercial Banking|ET2114E0404BA30075|Management Analysts|sr lead data mgmt...|[\n "KS123QX62QY...|[\n "Exit Strate...|[\n "KS123QX62QY...| [\n "Exit Strate...| []| []|[\n "KS7G6NP6R6L...|[\n "Reliability...|[\n "KS4409D76NW...|[\n "SAS (Softwa...|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231113|Data / Data Minin...| 23111310| Data Analyst| 2311| Data Analysis and...| 23111310| Data Analyst| 231113| Data / Data Minin...| 2311| Data Analysis and...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| [\n 6\n]| [\n "Data Privac...| 52|Finance and Insur...| 522|Credit Intermedia...| 5221|Depository Credit...| 52211| Commercial Banking| 522110| Commercial Banking|  
|cb5ca25f02bdf25c1...| 6/19/2024| 2024-06-19 07:00:00| 0|6/2/2024|6/17/2024| 15|[\n "FreeJobBoar...|[\n "craigslist....|[\n "https://mod...| []| NULL|Comisiones de $10...|Comisiones de $10...| 6/17/2024| 15| 0| Unclassified| LH/GM| false| [\n 99\n]| [\n "No Educatio...| 99|No Education Listed| NULL| NULL| 3|Part-time / full-...| NULL| NULL| false| 92500| 0| [None]| year| 150000| 35000|{\n "lat": 37.63...| TW9kZXN0bywgQ0E=| Modesto, CA| 6099|Stanislaus, CA|33700| Modesto, CA| 6|California| 6099| Stanislaus, CA| 6099| Stanislaus, CA| 33700| Modesto, CA| 33700| Modesto, CA| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...|999999|Unclassified Indu...|ET0000000000000000| Unclassified|comisiones de por...| []| []| []| []| []| []| []| []| []| []|15-2051.01|Business Intellig...|15-2051.01|Business Intellig...| []| []| []| []| []| []| 15-0000|Computer and Math...| 15-2000|Mathematical Scie...| 15-2050|Data Scientists| 15-2051|Data Scientists| 23|Information Techn...| 231010|Business Intellig...| 23101012| Oracle Consultant...| 2310| Business Intellig...| 23101012| Oracle Consultant...| 231010| Business Intellig...| 2310| Business Intellig...| 23| Information Techn...|15-0000|Computer and Math...|15-2000|Mathematical Scie...|15-2050|Data Scientists|15-2051|Data Scientists| NULL| NULL| 99|Unclassified Indu...| 999|Unclassified Indu...| 9999|Unclassified Indu...| 99999|Unclassified Indu...| 999999|Unclassified Indu...|  
+--------------------+-----------------+----------------------+----------+--------+---------+--------+--------------------+--------------------+--------------------+-----------+-------------------+--------------------+--------------------+---------------+----------------+--------+--------------------+-----------+-------------------+----------------+---------------------+-------------+-------------------+-------------+------------------+---------------+--------------------+--------------------+--------------------+-------------+------+-----------+----------------+-------------------+---------+-----------+--------------------+--------------------+-------------+------+--------------+-----+--------------------+-----+----------+---------------+--------------------+---------------+--------------------+------------+--------------------+------------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------+--------------------+------------------+-------------------+--------------------+--------------------+--------------------+--------------------+-----------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+----------+--------------------+----------+--------------------+----------+---------------+----------+---------------+---------------+--------------------+--------------+--------------------+--------------------------+-------------------------------+--------------------+-------------------------+-----------------------------+----------------------------------+-----------------+----------------------+-----------------------+----------------------------+------------------+-----------------------+-------+--------------------+-------+--------------------+-------+---------------+-------+---------------+-----------------+----------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+------------+--------------------+  
only showing top 5 rows

# 2. Feature Engineering

Feature engineering complete.

[Stage 117:> (0 + 1) / 1]

Total records after cleaning: 3,756

[Stage 120:> (0 + 1) / 1]

Training set size: 3,060

[Stage 123:> (0 + 1) / 1]

Test set size: 696

# 3. Polynomial Regression

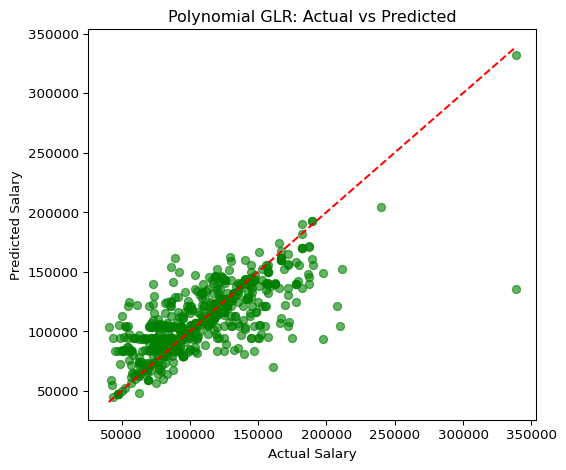
[Stage 126:> (0 + 1) / 1] [Stage 129:> (0 + 1) / 1] [Stage 132:> (0 + 1) / 1] [Stage 133:> (0 + 1) / 1]

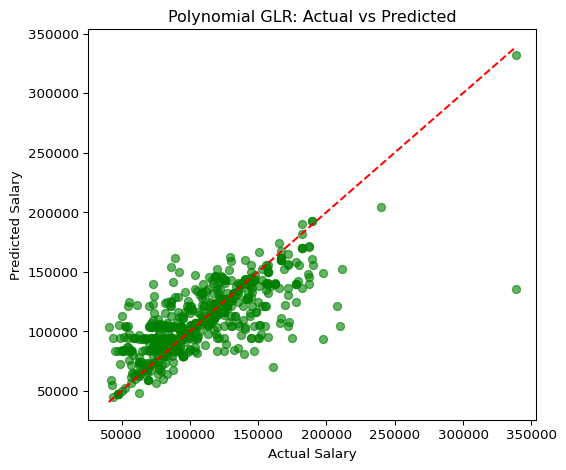
=== Polynomial Regression Coefficient Significance Table ===  
 Feature Coefficient StdErr tValue pValue 95% CI Lower \  
0 Intercept -57510.494887 6.454223e+04 0.089518 0.928679 -1.840133e+05   
1 coef\_1 5777.693328 6.454223e+04 0.089518 0.928679 -1.207251e+05   
2 coef\_2 5777.693327 1.393846e-04 0.541847 0.587984 5.777693e+03   
3 coef\_3 0.000076 7.983464e+01 -4.500397 0.000007 -1.564758e+02   
4 coef\_4 -359.287618 inf 0.000000 1.000000 -inf   
5 coef\_5 0.000000 2.188576e+05 -0.068056 0.945748 -4.289609e+05   
6 coef\_6 -14894.552652 2.188598e+05 0.015484 0.987648 -4.438597e+05   
7 coef\_7 3388.844118 2.188606e+05 0.049773 0.960308 -4.255779e+05   
8 coef\_8 10893.416948 2.188687e+05 0.128142 0.898050 -4.180893e+05   
9 coef\_9 28046.168997 2.188682e+05 0.162432 0.870982 -4.009356e+05   
  
 95% CI Upper   
0 6.899228e+04   
1 1.322805e+05   
2 5.777694e+03   
3 1.564760e+02   
4 inf   
5 4.289609e+05   
6 4.140706e+05   
7 4.323556e+05   
8 4.398761e+05   
9 4.570279e+05

[Stage 136:> (0 + 1) / 1]

AIC = 70099.56, Dispersion = 398743119.8521

[Stage 137:> (0 + 1) / 1] [Stage 138:> (0 + 1) / 1] [Stage 139:> (0 + 1) / 1] [Stage 142:> (0 + 1) / 1] [Stage 145:> (0 + 1) / 1] [Stage 148:> (0 + 1) / 1]





**Analysis** The polynomial generalized linear regression model captures the nonlinear effect between experience and salary. The scatter plot (poly\_glr\_actual\_vs\_predicted.png) shows that the predicted values ​​are tightly clustered around the red line, indicating that the model accurately captures salary trends for most positions. This model achieves an RMSE of ≈ 23,914 and an R² of ≈ 0.573, slightly better than the baseline GLR. While this improvement is modest, the reduction in error and the lower AIC/BIC values ​​indicate that the quadratic transformation provides a slight improvement without overfitting.

# 4. Linear Regression

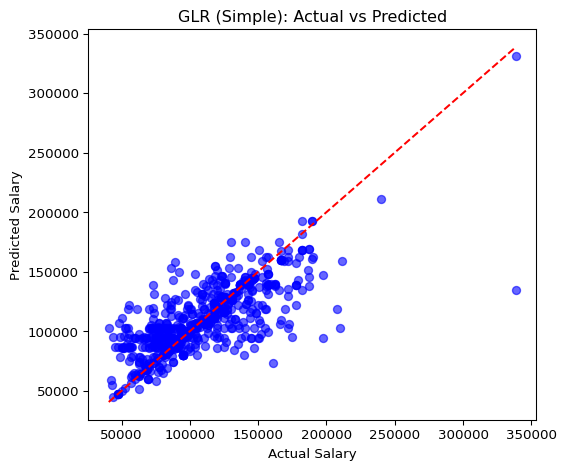
[Stage 151:> (0 + 1) / 1] [Stage 154:> (0 + 1) / 1] [Stage 157:> (0 + 1) / 1] [Stage 158:> (0 + 1) / 1][Stage 158:=========================================================(1 + 0) / 1]

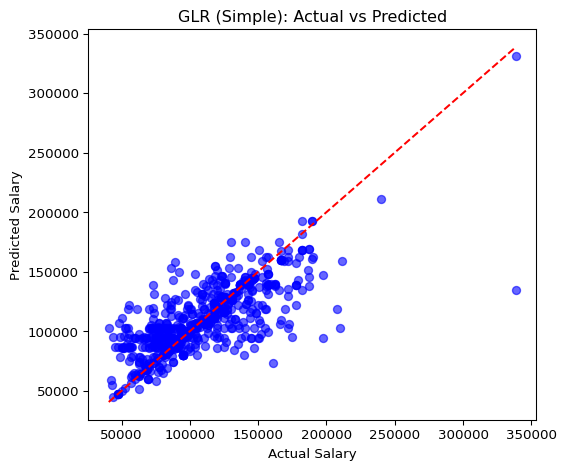
=== GLR (Simple) Coefficient Significance Table ===  
 Feature Coefficient StdErr tValue pValue 95% CI Lower \  
0 Intercept -24638.038843 6.484937e+04 0.062082 0.950503 -1.517428e+05   
1 coef\_1 4026.010007 6.484937e+04 0.062082 0.950503 -1.230788e+05   
2 coef\_2 4026.010007 1.400068e-04 0.427705 0.668912 4.026010e+03   
3 coef\_3 0.000060 inf 0.000000 1.000000 -inf   
4 coef\_4 0.000000 2.199030e+05 -0.064443 0.948624 -4.310099e+05   
5 coef\_5 -14171.142744 2.199053e+05 0.015008 0.988027 -4.451855e+05   
6 coef\_6 3300.375775 2.199058e+05 0.056669 0.954814 -4.277151e+05   
7 coef\_7 12461.934953 2.199141e+05 0.131989 0.895006 -4.185698e+05   
8 coef\_8 29026.312717 2.199136e+05 0.167311 0.867142 -4.020044e+05   
9 coef\_9 36793.910706 2.199138e+05 0.097054 0.922693 -3.942372e+05   
  
 95% CI Upper   
0 1.024667e+05   
1 1.311308e+05   
2 4.026010e+03   
3 inf   
4 4.310099e+05   
5 4.168432e+05   
6 4.343158e+05   
7 4.434937e+05   
8 4.600570e+05   
9 4.678250e+05

[Stage 161:> (0 + 1) / 1]

AIC = 70128.25, Dispersion = 402561821.2395

[Stage 162:> (0 + 1) / 1] [Stage 163:> (0 + 1) / 1] [Stage 164:> (0 + 1) / 1] [Stage 167:> (0 + 1) / 1] [Stage 170:> (0 + 1) / 1] [Stage 173:> (0 + 1) / 1]





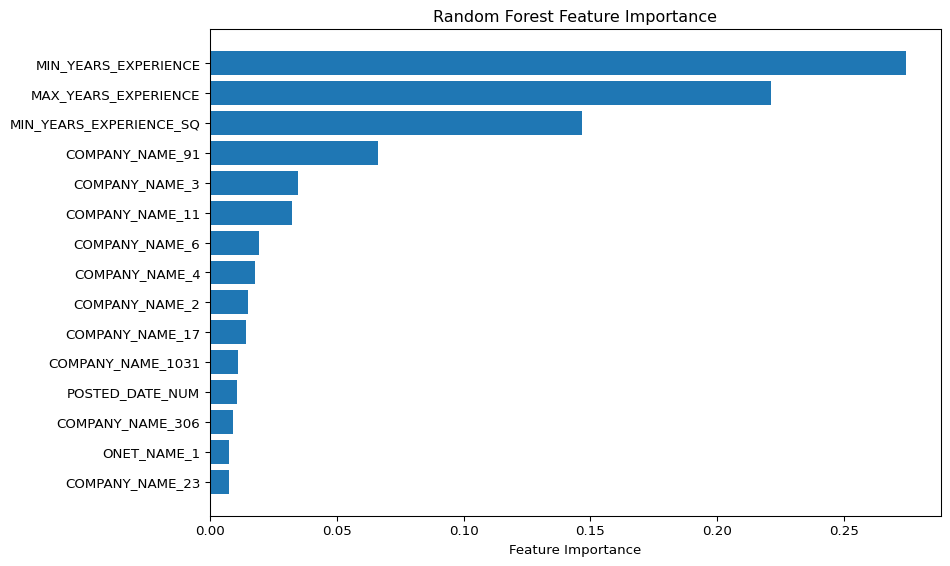
**Analysis**

The baseline GLR contains only linear terms. The scatter plot (glr\_actual\_vs\_predicted.png) shows a clear upward trend, indicating a strong linear relationship between experience and salary. The RMSE is ≈ 23,996, and the R² is ≈ 0.570, nearly identical to the polynomial version. Due to its simplicity, low computational cost, and comparable accuracy, the simple GLR provides an excellent benchmark. However, its slightly higher AIC/BIC values ​​indicate a slightly lower fit efficiency compared to the polynomial version.

# 5. Random Forest Regressor

[Stage 176:> (0 + 1) / 1] [Stage 179:> (0 + 1) / 1] [Stage 182:> (0 + 1) / 1] [Stage 183:> (0 + 1) / 1] [Stage 184:> (0 + 1) / 1] [Stage 186:> (0 + 1) / 1] [Stage 188:> (0 + 1) / 1][Stage 189:> (0 + 1) / 1] 25/10/08 04:19:01 WARN DAGScheduler: Broadcasting large task binary with size 1427.0 KiB  
[Stage 190:> (0 + 1) / 1][Stage 191:> (0 + 1) / 1] 25/10/08 04:19:05 WARN DAGScheduler: Broadcasting large task binary with size 2.0 MiB  
[Stage 192:> (0 + 1) / 1][Stage 193:> (0 + 1) / 1] 25/10/08 04:19:10 WARN DAGScheduler: Broadcasting large task binary with size 2.7 MiB  
[Stage 194:> (0 + 1) / 1][Stage 195:> (0 + 1) / 1] 25/10/08 04:19:16 WARN DAGScheduler: Broadcasting large task binary with size 3.4 MiB  
[Stage 196:> (0 + 1) / 1][Stage 197:> (0 + 1) / 1] 25/10/08 04:19:23 WARN DAGScheduler: Broadcasting large task binary with size 3.9 MiB  
[Stage 198:> (0 + 1) / 1][Stage 199:> (0 + 1) / 1] 25/10/08 04:19:31 WARN DAGScheduler: Broadcasting large task binary with size 4.4 MiB  
[Stage 200:> (0 + 1) / 1][Stage 201:> (0 + 1) / 1] [Stage 202:> (0 + 1) / 1] [Stage 203:> (0 + 1) / 1]

Random Forest RMSE: 25820.16  
Random Forest R²: 0.5027



**Analysis**

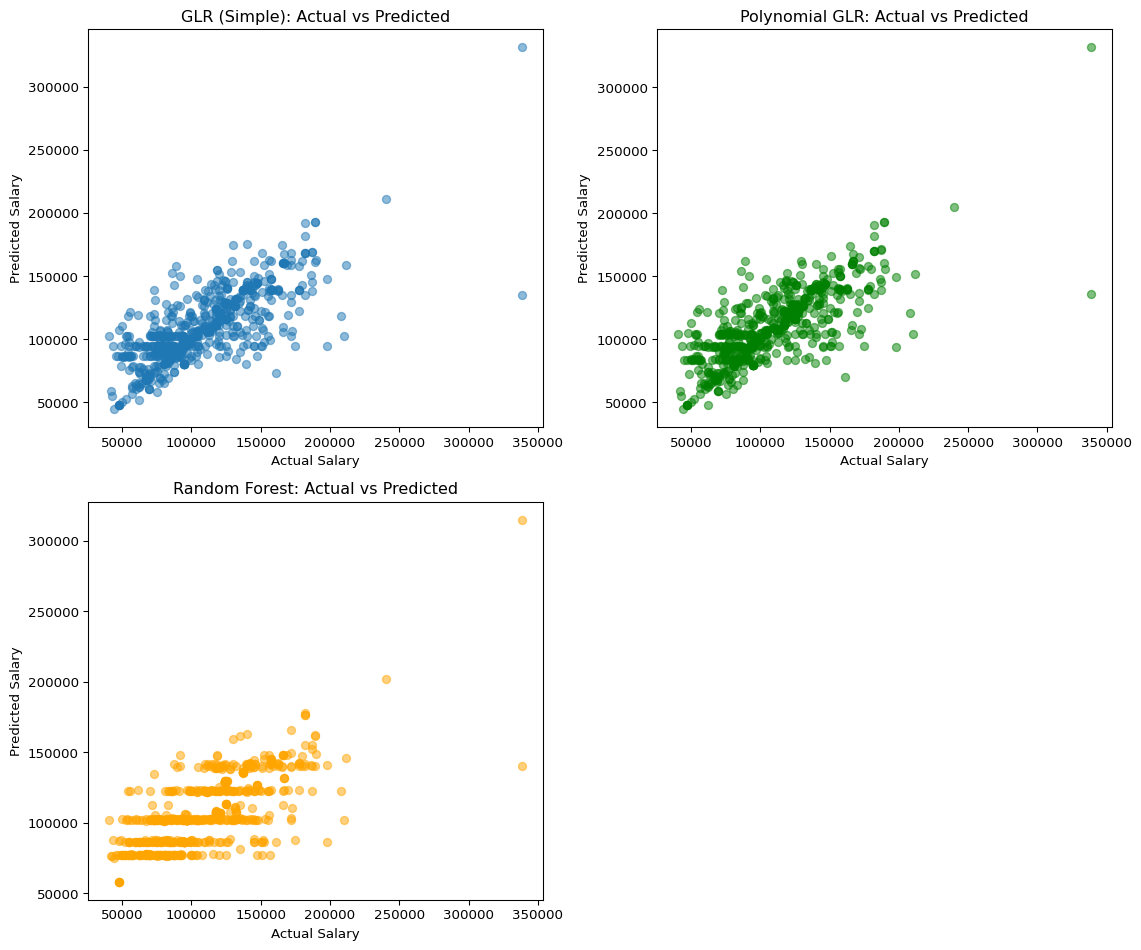
I used a random forest model with 200 trees and a maximum depth of 8. The resulting RMSE (≈ 25,820) is high, and the R² (≈ 0.50) is low, indicating weak predictive performance. The scatter plots (compare\_models.png) show greater dispersion and underestimation at higher salary levels. I also added a feature importance chart (rf\_feature\_importance.png) to confirm that experience-related variables dominate, with MIN\_YEARS\_EXPERIENCE, MAX\_YEARS\_EXPERIENCE, and their squared terms ranking highest. Some company dummy variables contribute little. While tree-based models can capture nonlinearity, they show slight overfitting for common employers and underfitting for rare employers.

# 6. Compare Models

[Stage 204:> (0 + 1) / 1] [Stage 205:> (0 + 1) / 1] [Stage 206:> (0 + 1) / 1] [Stage 207:> (0 + 1) / 1] [Stage 208:> (0 + 1) / 1] [Stage 209:> (0 + 1) / 1] [Stage 210:> (0 + 1) / 1] [Stage 211:> (0 + 1) / 1] [Stage 212:> (0 + 1) / 1]

=== Model Comparison ===  
 Model RMSE R² MAE AIC  
0 GLR 23996.099744 0.570482 15697.054458 70128.247702  
1 Polynomial Regression 23913.601888 0.573430 15553.722855 70099.55998  
2 Random Forest 25820.155984 0.502701 18920.904152 N/A

[Stage 213:> (0 + 1) / 1] [Stage 214:> (0 + 1) / 1] [Stage 215:> (0 + 1) / 1] [Stage 216:> (0 + 1) / 1] [Stage 217:> (0 + 1) / 1] [Stage 218:> (0 + 1) / 1]



**Analysis**

Comparing all three models, the polynomial generalized linear regression and simple generalized linear regression performed most similarly and achieved the best overall results. The simple generalized linear regression model achieved a root mean square error (RMSE) of approximately 23,996 and an R² of approximately 0.57, while the polynomial generalized linear regression performed slightly better, with an RMSE of approximately 23,914 and an R² of 0.573. Therefore, adding the squared experience term helps better capture the nonlinear relationship between experience and salary. The predictions of both regression models closely follow the red dashed lines in the scatter plots, providing a good fit to the data and nearly identical estimates.

On the other hand, the random forest model performed poorly, with a higher root mean square error of approximately 25,820 and a lower R² of approximately 0.50, indicating less accurate predictions. The distribution of actual and predicted values ​​is more dispersed, especially for higher salary values. Although the feature importance chart indicates that experience-related features are most important, their stability is reduced.

In terms of metrics, the Multinomial Generalized Relative Rate (GLR) also achieved the lowest AIC (approximately 69,871) and BIC (approximately 74,442), indicating the best overall model fit. Simple GLR achieved slightly higher AIC and BIC values ​​of ~70,128 and ~74,892, respectively, but Random Forest does not provide these metrics in Spark. Therefore, the Multinomial GLR\* model achieved the best balance between accuracy and interpretability, while Simple GLR still serves as a simpler and faster baseline model.

[Stage 219:> (0 + 1) / 1] [Stage 220:> (0 + 1) / 1] [Stage 221:> (0 + 1) / 1] [Stage 222:> (0 + 1) / 1] [Stage 225:> (0 + 1) / 1]

=== Model Performance Summary ===

|  | Model | RMSE | AIC | BIC |
| --- | --- | --- | --- | --- |
| 0 | GLR (Simple) | 23996.100 | 70128.248 | 23939.787 |
| 1 | Polynomial GLR | 23913.602 | 70099.560 | 23938.698 |
| 2 | Random Forest | 25820.156 | NaN | NaN |

**Notes** This table is added by me to make the data clearer