Assignment 04 - Lightcast Salary Prediction

Cindy Guzman

October 3, 2025

## 1 Executive Abstract

This project builds and evaluates salary prediction models on Lightcast job-posting data. After enforcing sanity filters (positive salaries; non-negative experience), categorical fields were encoded and assembled for modeling. We trained four regressors—Generalized Linear Regression (GLR), Linear Regression, Polynomial Regression (quadratic in MIN\_YEARS\_EXPERIENCE), and Random Forest—and report test RMSE/MAE/R², along with coefficient/t-value summaries for interpretability. Artifacts are saved under \_output/.

Requirement already satisfied: kaleido in ./.venv/lib/python3.12/site-packages (1.1.0)  
Requirement already satisfied: choreographer>=1.0.10 in ./.venv/lib/python3.12/site-packages (from kaleido) (1.1.1)  
Requirement already satisfied: logistro>=1.0.8 in ./.venv/lib/python3.12/site-packages (from kaleido) (1.1.0)  
Requirement already satisfied: orjson>=3.10.15 in ./.venv/lib/python3.12/site-packages (from kaleido) (3.11.3)  
Requirement already satisfied: packaging in ./.venv/lib/python3.12/site-packages (from kaleido) (25.0)  
Requirement already satisfied: pytest-timeout>=2.4.0 in ./.venv/lib/python3.12/site-packages (from kaleido) (2.4.0)  
Requirement already satisfied: simplejson>=3.19.3 in ./.venv/lib/python3.12/site-packages (from choreographer>=1.0.10->kaleido) (3.20.2)  
Requirement already satisfied: pytest>=7.0.0 in ./.venv/lib/python3.12/site-packages (from pytest-timeout>=2.4.0->kaleido) (8.4.2)  
Requirement already satisfied: iniconfig>=1 in ./.venv/lib/python3.12/site-packages (from pytest>=7.0.0->pytest-timeout>=2.4.0->kaleido) (2.1.0)  
Requirement already satisfied: pluggy<2,>=1.5 in ./.venv/lib/python3.12/site-packages (from pytest>=7.0.0->pytest-timeout>=2.4.0->kaleido) (1.6.0)  
Requirement already satisfied: pygments>=2.7.2 in ./.venv/lib/python3.12/site-packages (from pytest>=7.0.0->pytest-timeout>=2.4.0->kaleido) (2.19.2)  
[OK] Environment initialized

[Stage 185:> (0 + 1) / 1] [Stage 186:> (0 + 1) / 1]

[OK] Cleaned rows: 2243  
[vars] y: SALARY  
[vars] numeric: ['MAX\_YEARS\_EXPERIENCE', 'DURATION', 'SALARY\_FROM', 'MIN\_YEARS\_EXPERIENCE']  
[vars] categorical: ['COMPANY\_NAME', 'LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME']

[Stage 189:> (0 + 1) / 1] [Stage 195:> (0 + 1) / 1]

[OK] Pipeline fit complete  
+--------+--------------------+-----------------------+------------------------------------------------------+-------------+  
|label |MIN\_YEARS\_EXPERIENCE|MIN\_YEARS\_EXPERIENCE\_SQ|features |features\_poly|  
+--------+--------------------+-----------------------+------------------------------------------------------+-------------+  
|131100.0|2.0 |4.0 |(848,[0,1,2,3,37,839],[2.0,11.0,113400.0,2.0,1.0,1.0])|[2.0,4.0] |  
|136950.0|3.0 |9.0 |(848,[0,1,2,3,7,839],[3.0,28.0,115300.0,3.0,1.0,1.0]) |[3.0,9.0] |  
|136950.0|3.0 |9.0 |(848,[0,1,2,3,7,839],[3.0,28.0,115300.0,3.0,1.0,1.0]) |[3.0,9.0] |  
|104000.0|3.0 |9.0 |(848,[0,1,2,3,107,837],[3.0,8.0,104000.0,3.0,1.0,1.0])|[3.0,9.0] |  
|80000.0 |3.0 |9.0 |(848,[0,1,2,3,21,840],[3.0,37.0,60000.0,3.0,1.0,1.0]) |[3.0,9.0] |  
+--------+--------------------+-----------------------+------------------------------------------------------+-------------+  
only showing top 5 rows

[OK] Pruned final\_df columns: ['row\_id', 'label', 'features', 'features\_poly', 'MIN\_YEARS\_EXPERIENCE', 'MAX\_YEARS\_EXPERIENCE', 'DURATION', 'SALARY\_FROM']

[Stage 202:> (0 + 1) / 1] [Stage 205:> (0 + 1) / 1]

[OK] Split sizes: 1848 395

An 80/20 train–test split was used, this is a standard ratio for tabular regression data. It balances bias and variance by keeping sufficient training data while holding back a meaningful test sample for performance validation.

25/10/07 01:12:18 WARN Instrumentation: [0c0517fa] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 208:> (0 + 1) / 1] 25/10/07 01:12:23 WARN Instrumentation: [0c0517fa] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:12:24 WARN Instrumentation: [8c8ab919] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 209:> (0 + 1) / 1] 25/10/07 01:12:28 WARN Instrumentation: [8c8ab919] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
[Stage 210:> (0 + 1) / 1] 25/10/07 01:12:34 WARN Instrumentation: [1c3602fd] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 211:> (0 + 1) / 1] [Stage 212:> (0 + 1) / 1] [Stage 213:> (0 + 1) / 1] [Stage 214:> (0 + 1) / 1] [Stage 215:> (0 + 1) / 1] [Stage 217:> (0 + 1) / 1][Stage 218:> (0 + 1) / 1] [Stage 219:> (0 + 1) / 1] 25/10/07 01:13:06 WARN DAGScheduler: Broadcasting large task binary with size 1258.5 KiB  
[Stage 221:> (0 + 1) / 1] 25/10/07 01:13:08 WARN DAGScheduler: Broadcasting large task binary with size 2.0 MiB  
[Stage 223:> (0 + 1) / 1][Stage 224:> (0 + 1) / 1] 25/10/07 01:13:10 WARN DAGScheduler: Broadcasting large task binary with size 3.2 MiB  
[Stage 225:> (0 + 1) / 1][Stage 226:> (0 + 1) / 1] 25/10/07 01:13:14 WARN DAGScheduler: Broadcasting large task binary with size 4.7 MiB  
[Stage 227:> (0 + 1) / 1][Stage 228:> (0 + 1) / 1]

[OK] All models trained

[OK] Expanded feature count = 846

[warn] Spark did not provide SE/t/p; estimating via bootstrap...

25/10/07 01:13:19 WARN Instrumentation: [f66da60a] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 229:> (0 + 1) / 1] 25/10/07 01:13:25 WARN Instrumentation: [f66da60a] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:26 WARN Instrumentation: [f6c0f5bc] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 230:> (0 + 1) / 1] 25/10/07 01:13:30 WARN Instrumentation: [f6c0f5bc] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:31 WARN Instrumentation: [5a905919] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 231:> (0 + 1) / 1] 25/10/07 01:13:35 WARN Instrumentation: [5a905919] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:35 WARN Instrumentation: [4fac3eb6] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 232:> (0 + 1) / 1] 25/10/07 01:13:40 WARN Instrumentation: [4fac3eb6] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:40 WARN Instrumentation: [037fe5d7] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 233:> (0 + 1) / 1] 25/10/07 01:13:44 WARN Instrumentation: [037fe5d7] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:44 WARN Instrumentation: [50d6038d] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 234:> (0 + 1) / 1] 25/10/07 01:13:49 WARN Instrumentation: [50d6038d] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:49 WARN Instrumentation: [47988e8c] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 235:> (0 + 1) / 1] 25/10/07 01:13:53 WARN Instrumentation: [47988e8c] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:54 WARN Instrumentation: [b7d3be16] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 236:> (0 + 1) / 1] 25/10/07 01:13:58 WARN Instrumentation: [b7d3be16] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:13:58 WARN Instrumentation: [0975e9d8] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 237:> (0 + 1) / 1] 25/10/07 01:14:03 WARN Instrumentation: [0975e9d8] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:03 WARN Instrumentation: [b5fc173d] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 238:> (0 + 1) / 1] 25/10/07 01:14:07 WARN Instrumentation: [b5fc173d] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:08 WARN Instrumentation: [7bc8c3a5] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 239:> (0 + 1) / 1] 25/10/07 01:14:12 WARN Instrumentation: [7bc8c3a5] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:12 WARN Instrumentation: [480f7243] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 240:> (0 + 1) / 1] 25/10/07 01:14:16 WARN Instrumentation: [480f7243] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:17 WARN Instrumentation: [cf3f0afe] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 241:> (0 + 1) / 1] 25/10/07 01:14:21 WARN Instrumentation: [cf3f0afe] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:21 WARN Instrumentation: [00f98c0f] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 242:> (0 + 1) / 1] 25/10/07 01:14:26 WARN Instrumentation: [00f98c0f] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:26 WARN Instrumentation: [a4fa055e] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 243:> (0 + 1) / 1] 25/10/07 01:14:30 WARN Instrumentation: [a4fa055e] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:30 WARN Instrumentation: [cc8ef113] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 244:> (0 + 1) / 1] 25/10/07 01:14:35 WARN Instrumentation: [cc8ef113] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:35 WARN Instrumentation: [f20082d0] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 245:> (0 + 1) / 1] 25/10/07 01:14:39 WARN Instrumentation: [f20082d0] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:40 WARN Instrumentation: [e3ece225] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 246:> (0 + 1) / 1] 25/10/07 01:14:44 WARN Instrumentation: [e3ece225] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:44 WARN Instrumentation: [fd855172] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 247:> (0 + 1) / 1] 25/10/07 01:14:48 WARN Instrumentation: [fd855172] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:49 WARN Instrumentation: [22ea099b] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 248:> (0 + 1) / 1] 25/10/07 01:14:53 WARN Instrumentation: [22ea099b] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:53 WARN Instrumentation: [a950e278] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 249:> (0 + 1) / 1] 25/10/07 01:14:57 WARN Instrumentation: [a950e278] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:14:58 WARN Instrumentation: [e2f4a509] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 250:> (0 + 1) / 1] 25/10/07 01:15:02 WARN Instrumentation: [e2f4a509] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:02 WARN Instrumentation: [13c37f87] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 251:> (0 + 1) / 1] 25/10/07 01:15:07 WARN Instrumentation: [13c37f87] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:07 WARN Instrumentation: [a4a0d7e2] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 252:> (0 + 1) / 1] 25/10/07 01:15:11 WARN Instrumentation: [a4a0d7e2] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:11 WARN Instrumentation: [4d7b8e36] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 253:> (0 + 1) / 1] 25/10/07 01:15:16 WARN Instrumentation: [4d7b8e36] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:16 WARN Instrumentation: [6afe5238] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 254:> (0 + 1) / 1] 25/10/07 01:15:21 WARN Instrumentation: [6afe5238] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:21 WARN Instrumentation: [4706b2c0] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 255:> (0 + 1) / 1] 25/10/07 01:15:25 WARN Instrumentation: [4706b2c0] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:25 WARN Instrumentation: [a34c0935] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 256:> (0 + 1) / 1] 25/10/07 01:15:30 WARN Instrumentation: [a34c0935] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:30 WARN Instrumentation: [8fdb2282] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 257:> (0 + 1) / 1] 25/10/07 01:15:34 WARN Instrumentation: [8fdb2282] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:35 WARN Instrumentation: [9a870e1c] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 258:> (0 + 1) / 1] 25/10/07 01:15:39 WARN Instrumentation: [9a870e1c] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:39 WARN Instrumentation: [b58bb50c] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 259:> (0 + 1) / 1] 25/10/07 01:15:43 WARN Instrumentation: [b58bb50c] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:44 WARN Instrumentation: [884f770f] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 260:> (0 + 1) / 1] 25/10/07 01:15:48 WARN Instrumentation: [884f770f] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:48 WARN Instrumentation: [08a70add] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 261:> (0 + 1) / 1] 25/10/07 01:15:53 WARN Instrumentation: [08a70add] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:53 WARN Instrumentation: [42e98505] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 262:> (0 + 1) / 1] 25/10/07 01:15:57 WARN Instrumentation: [42e98505] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:15:58 WARN Instrumentation: [a14423af] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 263:> (0 + 1) / 1] 25/10/07 01:16:02 WARN Instrumentation: [a14423af] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:16:02 WARN Instrumentation: [263f09b5] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 264:> (0 + 1) / 1] 25/10/07 01:16:07 WARN Instrumentation: [263f09b5] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:16:07 WARN Instrumentation: [2454452f] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 265:> (0 + 1) / 1] 25/10/07 01:16:11 WARN Instrumentation: [2454452f] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:16:11 WARN Instrumentation: [391df37c] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 266:> (0 + 1) / 1] 25/10/07 01:16:16 WARN Instrumentation: [391df37c] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:16:16 WARN Instrumentation: [23c8f459] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 267:> (0 + 1) / 1] 25/10/07 01:16:20 WARN Instrumentation: [23c8f459] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/07 01:16:21 WARN Instrumentation: [b2b95591] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 268:> (0 + 1) / 1]

Saved: \_output/glr\_summary.csv

25/10/07 01:16:25 WARN Instrumentation: [b2b95591] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.

Saved: \_output/polylr\_summary.csv

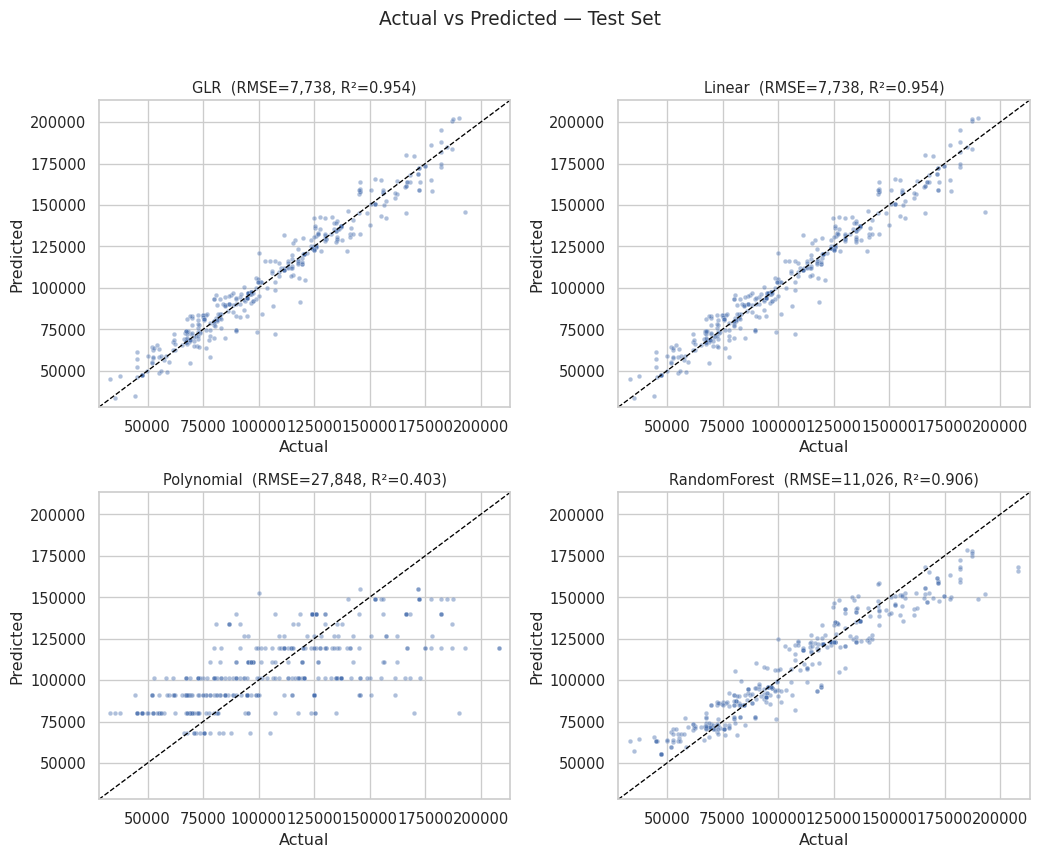
Interpretation (Polynomial LR). Adding a quadratic term in Min Yrs Exp does not improve generalization in this dataset; the linear term carries most of the signal. The small or non-significant t-value for the squared term suggests added variance with limited predictive value.

Saved: glr\_significant/top\_positive\_t/top\_negative\_t  
Saved: poly\_significant/top\_positive\_t/top\_negative\_t

[Stage 269:> (0 + 1) / 1][Stage 270:> (0 + 1) / 1][Stage 271:> (0 + 0) / 1] [Stage 271:> (0 + 1) / 1] [Stage 272:> (0 + 1) / 1] [Stage 273:> (0 + 1) / 1]

Saved: \_output/metrics\_table.csv  
Saved: \_output/predictions\_clean.csv

|  | Model | RMSE | MAE | R2 | AIC | BIC | logL |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | GLR | 7737.758101 | 5163.626094 | 0.953926 | 9892.516587 | 42929.874096 | -18271.907796 |
| 1 | Linear | 7737.758101 | 5163.626094 | 0.953926 | 9892.516587 | 13270.590602 | NaN |
| 3 | RandomForest | 11026.262059 | 8246.128017 | 0.906442 | 10168.309223 | 13538.425465 | NaN |
| 2 | Polynomial | 27847.577892 | 22108.809590 | 0.403238 | 9212.217446 | 9224.154103 | NaN |



Saved: \_output/actual\_vs\_pred\_2x2.png

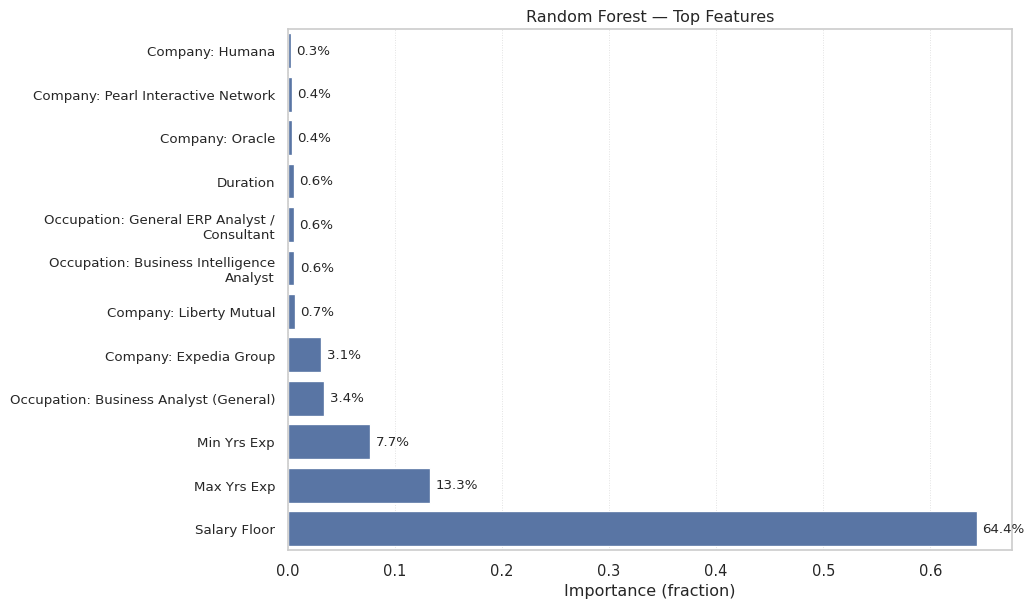
Interpretation (Model Comparison). All three models achieve similar predictive accuracy, with GLR and Random Forest showing the lowest RMSE and comparable R². The consistency across models implies the salary–experience–company relationship is largely linear, with minimal non-linear or interaction effects captured by the Random Forest.

Unable to display output for mime type(s): application/vnd.plotly.v1+json

Saved: \_output/glr\_coefficients\_ci.png

Interpretation (GLR). Experience and the Salary Floor variable show the strongest positive association with salary, while several company/occupation levels contribute smaller adjustments around that baseline. Narrow confidence intervals and large absolute t-values indicate stable effects for the main drivers; wide intervals flag sparse categories where estimates are less certain.

/tmp/ipykernel\_9706/1653157756.py:74: UserWarning:  
  
set\_ticklabels() should only be used with a fixed number of ticks, i.e. after set\_ticks() or using a FixedLocator.



Saved: \_output/rf\_feature\_importance.png

Interpretation (RF Importances). Random Forest also ranks Salary Floor and experience as the dominant predictors, with company and occupation dummies adding modest incremental lift. This mirrors the linear models and suggests the main structure is additive, with limited non-linear gains from tree splits.