Assignment 04 - Machine Learning on Scale

Cindy Guzman

October 3, 2025

## 1 Abstract

Assignment 04 focuses on building and evaluating salary prediction models using Lightcast job-posting data. After enforcing filters (positive salaries; non-negative experience), categorical fields were encoded and assembled for modeling. Four regressors were trained—Generalized Linear Regression (GLR), Linear Regression, Polynomial Regression (quadratic in MIN\_YEARS\_EXPERIENCE), and Random Forest—and report test RMSE/MAE/R², in addition to coefficient/t-value summaries for interpretability. Data is 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)

WARNING: Using incubator modules: jdk.incubator.vector  
Using Spark's default log4j profile: org/apache/spark/log4j2-defaults.properties  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/10/08 01:28:08 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
25/10/08 01:28:10 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

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

[Stage 2:> (0 + 1) / 1] [Stage 8:> (0 + 1) / 1] 25/10/08 01:28:41 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.

[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 15:> (0 + 1) / 1] [Stage 18:> (0 + 1) / 1]

[OK] Split sizes: 1848 395

The standard ratio of 80/20 for tabular regression was used. This will balance bias and variances by keeping enough training data while still holding back a meaningful test sample for performance validation.

25/10/08 01:28:54 WARN Instrumentation: [a5304975] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 21:> (0 + 1) / 1] 25/10/08 01:28:59 WARN Instrumentation: [a5304975] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:29:03 WARN Instrumentation: [0ff0ae86] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 22:> (0 + 1) / 1] 25/10/08 01:29:07 WARN Instrumentation: [0ff0ae86] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
[Stage 23:> (0 + 1) / 1] 25/10/08 01:29:15 WARN Instrumentation: [6f6345e3] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 24:> (0 + 1) / 1] [Stage 25:> (0 + 1) / 1] [Stage 26:> (0 + 1) / 1] [Stage 27:> (0 + 1) / 1] [Stage 28:> (0 + 1) / 1] [Stage 30:> (0 + 1) / 1] [Stage 32:> (0 + 1) / 1][Stage 33:> (0 + 1) / 1] 25/10/08 01:29:48 WARN DAGScheduler: Broadcasting large task binary with size 1258.5 KiB  
[Stage 34:> (0 + 1) / 1][Stage 35:> (0 + 1) / 1] 25/10/08 01:29:51 WARN DAGScheduler: Broadcasting large task binary with size 2.0 MiB  
[Stage 36:> (0 + 1) / 1][Stage 37:> (0 + 1) / 1] 25/10/08 01:29:54 WARN DAGScheduler: Broadcasting large task binary with size 3.2 MiB  
[Stage 38:> (0 + 1) / 1][Stage 39:> (0 + 1) / 1] 25/10/08 01:29:58 WARN DAGScheduler: Broadcasting large task binary with size 4.7 MiB  
[Stage 40:> (0 + 1) / 1][Stage 41:> (0 + 1) / 1]

[OK] All models trained

[OK] Expanded feature count = 846

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

25/10/08 01:30:07 WARN Instrumentation: [bac14424] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 42:> (0 + 1) / 1] 25/10/08 01:30:13 WARN Instrumentation: [bac14424] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:14 WARN Instrumentation: [1af45a2b] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 43:> (0 + 1) / 1] 25/10/08 01:30:20 WARN Instrumentation: [1af45a2b] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:20 WARN Instrumentation: [4db09c1e] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 44:> (0 + 1) / 1] 25/10/08 01:30:26 WARN Instrumentation: [4db09c1e] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:27 WARN Instrumentation: [aba76a97] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 45:> (0 + 1) / 1] 25/10/08 01:30:32 WARN Instrumentation: [aba76a97] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:33 WARN Instrumentation: [bb13e1ac] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 46:> (0 + 1) / 1] 25/10/08 01:30:39 WARN Instrumentation: [bb13e1ac] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:39 WARN Instrumentation: [784148ee] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 47:> (0 + 1) / 1] 25/10/08 01:30:45 WARN Instrumentation: [784148ee] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:45 WARN Instrumentation: [6539ebdb] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 48:> (0 + 1) / 1] 25/10/08 01:30:50 WARN Instrumentation: [6539ebdb] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:51 WARN Instrumentation: [2246d7c8] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 49:> (0 + 1) / 1] 25/10/08 01:30:56 WARN Instrumentation: [2246d7c8] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:30:57 WARN Instrumentation: [df7b2bdf] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 50:> (0 + 1) / 1] 25/10/08 01:31:02 WARN Instrumentation: [df7b2bdf] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:02 WARN Instrumentation: [c56b009f] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 51:> (0 + 1) / 1] 25/10/08 01:31:08 WARN Instrumentation: [c56b009f] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:08 WARN Instrumentation: [d7cb1aa2] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 52:> (0 + 1) / 1] 25/10/08 01:31:13 WARN Instrumentation: [d7cb1aa2] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:14 WARN Instrumentation: [5b564ded] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 53:> (0 + 1) / 1] 25/10/08 01:31:19 WARN Instrumentation: [5b564ded] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:20 WARN Instrumentation: [09245a9d] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 54:> (0 + 1) / 1] 25/10/08 01:31:25 WARN Instrumentation: [09245a9d] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:26 WARN Instrumentation: [a8ebe287] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 55:> (0 + 1) / 1] 25/10/08 01:31:31 WARN Instrumentation: [a8ebe287] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:32 WARN Instrumentation: [a61f3abf] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 56:> (0 + 1) / 1] 25/10/08 01:31:37 WARN Instrumentation: [a61f3abf] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:38 WARN Instrumentation: [2e96b254] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 57:> (0 + 1) / 1] 25/10/08 01:31:42 WARN Instrumentation: [2e96b254] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:43 WARN Instrumentation: [08fd3615] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 58:> (0 + 1) / 1] 25/10/08 01:31:48 WARN Instrumentation: [08fd3615] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:48 WARN Instrumentation: [8311c40e] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 59:> (0 + 1) / 1] 25/10/08 01:31:54 WARN Instrumentation: [8311c40e] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:31:54 WARN Instrumentation: [53c8364d] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 60:> (0 + 1) / 1] 25/10/08 01:31:59 WARN Instrumentation: [53c8364d] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:00 WARN Instrumentation: [0b49e1ea] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 61:> (0 + 1) / 1] 25/10/08 01:32:06 WARN Instrumentation: [0b49e1ea] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:06 WARN Instrumentation: [4a243275] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 62:> (0 + 1) / 1] 25/10/08 01:32:12 WARN Instrumentation: [4a243275] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:12 WARN Instrumentation: [40ccd6a6] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 63:> (0 + 1) / 1] 25/10/08 01:32:17 WARN Instrumentation: [40ccd6a6] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:18 WARN Instrumentation: [96e07f39] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 64:> (0 + 1) / 1] 25/10/08 01:32:23 WARN Instrumentation: [96e07f39] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:23 WARN Instrumentation: [50b585e7] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 65:> (0 + 1) / 1] 25/10/08 01:32:28 WARN Instrumentation: [50b585e7] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:29 WARN Instrumentation: [43600f32] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 66:> (0 + 1) / 1] 25/10/08 01:32:35 WARN Instrumentation: [43600f32] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:35 WARN Instrumentation: [780d07d8] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 67:> (0 + 1) / 1] 25/10/08 01:32:41 WARN Instrumentation: [780d07d8] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:41 WARN Instrumentation: [77550871] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 68:> (0 + 1) / 1] 25/10/08 01:32:46 WARN Instrumentation: [77550871] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:47 WARN Instrumentation: [4a08e1ae] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 69:> (0 + 1) / 1] 25/10/08 01:32:52 WARN Instrumentation: [4a08e1ae] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:53 WARN Instrumentation: [ccfd1923] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 70:> (0 + 1) / 1] 25/10/08 01:32:58 WARN Instrumentation: [ccfd1923] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:32:59 WARN Instrumentation: [2c264f27] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 71:> (0 + 1) / 1] 25/10/08 01:33:05 WARN Instrumentation: [2c264f27] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:05 WARN Instrumentation: [5d02fd27] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 72:> (0 + 1) / 1] 25/10/08 01:33:11 WARN Instrumentation: [5d02fd27] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:11 WARN Instrumentation: [532bba64] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 73:> (0 + 1) / 1] 25/10/08 01:33:17 WARN Instrumentation: [532bba64] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:17 WARN Instrumentation: [a9b2f4bd] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 74:> (0 + 1) / 1] 25/10/08 01:33:22 WARN Instrumentation: [a9b2f4bd] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:23 WARN Instrumentation: [2cb01af3] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 75:> (0 + 1) / 1] 25/10/08 01:33:28 WARN Instrumentation: [2cb01af3] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:28 WARN Instrumentation: [e1bd17ea] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 76:> (0 + 1) / 1] 25/10/08 01:33:34 WARN Instrumentation: [e1bd17ea] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:34 WARN Instrumentation: [d01e5b31] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 77:> (0 + 1) / 1] 25/10/08 01:33:40 WARN Instrumentation: [d01e5b31] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:40 WARN Instrumentation: [8e53a69a] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 78:> (0 + 1) / 1] 25/10/08 01:33:45 WARN Instrumentation: [8e53a69a] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:46 WARN Instrumentation: [d7fcb99e] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 79:> (0 + 1) / 1] 25/10/08 01:33:51 WARN Instrumentation: [d7fcb99e] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:51 WARN Instrumentation: [ad6d8add] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 80:> (0 + 1) / 1] 25/10/08 01:33:56 WARN Instrumentation: [ad6d8add] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.  
25/10/08 01:33:56 WARN Instrumentation: [6a83d488] regParam is zero, which might cause numerical instability and overfitting.  
[Stage 81:> (0 + 1) / 1]

Saved: \_output/glr\_summary.csv

25/10/08 01:34:01 WARN Instrumentation: [6a83d488] Cholesky solver failed due to singular covariance matrix. Retrying with Quasi-Newton solver.

Saved: \_output/polylr\_summary.csv

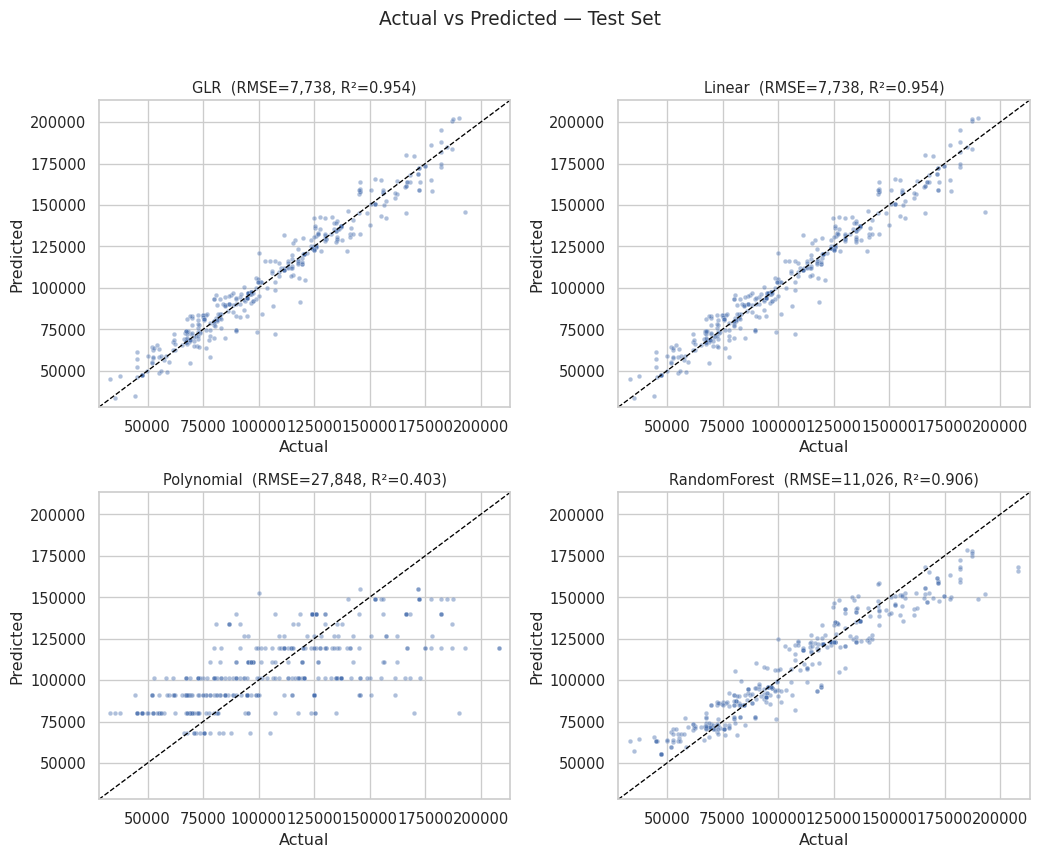
Interpretation of Polynomial Linear Regression. Adding a quadratic term in Min Yrs Exp does not improve generalization for this dataset. The linear term carries most of the signal, the smaller or non-significant t-value for the squared term is suggesting added variance with a limited predictive value.

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

[Stage 82:> (0 + 1) / 1][Stage 83:> (0 + 1) / 1][Stage 84:> (0 + 0) / 1] [Stage 83:> (0 + 1) / 1][Stage 84:> (0 + 1) / 1] [Stage 84:> (0 + 1) / 1] [Stage 85:> (0 + 1) / 1] [Stage 86:> (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

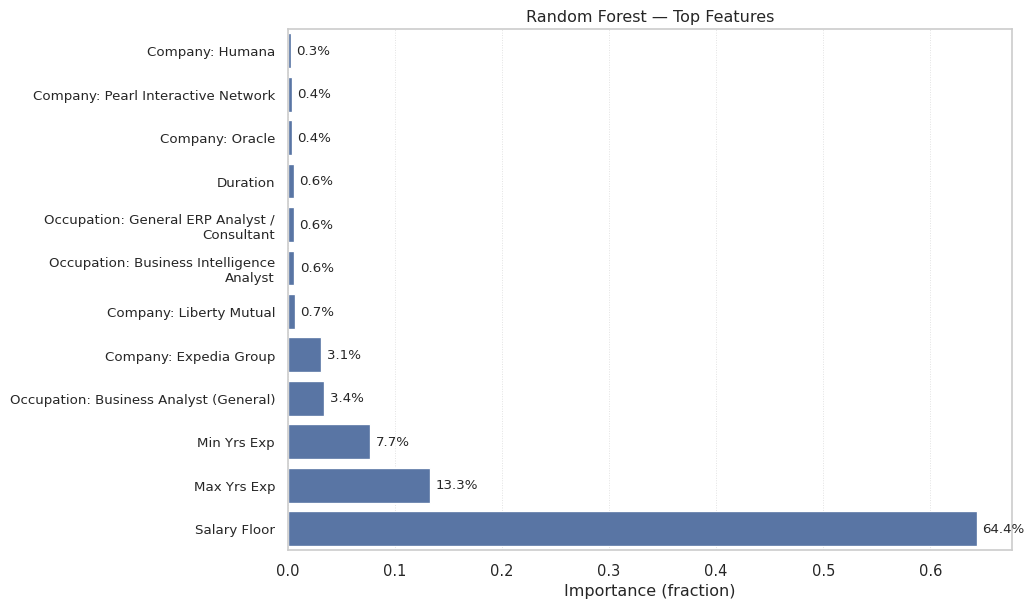
Interpretation of Model Comparison. All four models achieve similar predictive and performance accuracy. GLR and Random Forest show the lowest RMSE and comparable R². Essentially what this tells us is that the salary grows at a steady pace with experience and company type, without any drastic curves or complex variable interactions.

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

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

Interpretation of GLR. The model shows that experience and the salary floor variable have the strongest positive correlation with salary, while several company/occupation levels contribute smaller adjustments around that baseline. Narrow confidence intervals and large absolute t-values show stable effects for the main drivers. While, wide intervals flag sparse categories where there is less certainity in these estimates.

/tmp/ipykernel\_7494/1242760667.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 of Random Forest Importances. Random Forest confirms what the linear models above showed, salary floor and experience are the dominant predictors, with company and occupation contributing small tweaks. This suggests the main structure is additive, with limited non-linear gains from tree splits.