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/R^2, 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 02:13:11 WARN NativeCodeLoader: Unable to load native-hadoop library for your platfo
java classes where applicable
25/10/08 02:13:12 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port

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

[Stage 2:>                                                          (0 + 1) / 1]

[OK] Pipeline fit complete
+--------+--------------------+----------------------+-----------------------------------
+-------------+
|label   |MIN_YEARS_EXPERIENCE|MIN_YEARS_EXPERIENCE_SQ|features
+--------+--------------------+----------------------+-----------------------------------
+-------------+
|131100.0|2.0                 |4.0                   |(848,[0,1,2,3,37,839],[2.0,11.0,11340
|136950.0|3.0                 |9.0                   |(848,[0,1,2,3,7,839],[3.0,28.0,115300
|136950.0|3.0                 |9.0                   |(848,[0,1,2,3,7,839],[3.0,28.0,115300
|104000.0|3.0                 |9.0                   |(848,[0,1,2,3,107,837],[3.0,8.0,10400
|80000.0 |3.0                 |9.0                   |(848,[0,1,2,3,21,840],[3.0,37.0,60000
+--------+--------------------+----------------------+-----------------------------------
+-------------+
only showing top 5 rows

[OK] Pruned final_df columns: ['row_id', 'label', 'features', 'features_poly', 'MIN_YEARS_EXI

[Stage 15:>                                                         (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 02:13:55 WARN Instrumentation: [157afd70] regParam is zero, which might cause numeri
[Stage 21:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:14:04 WARN Instrumentation: [5e265ea5] regParam is zero, which might cause numeri
[Stage 22:>                                                         (0 + 1) / 1]
Newton solver.
[Stage 23:>                                                         (0 + 1) / 1]
[Stage 24:>                                                         (0 + 1) / 1]
[Stage 34:>                                                         (0 + 1) / 1][Stage 35:>
[Stage 36:>                                                         (0 + 1) / 1][Stage 37:>
[Stage 38:>                                                         (0 + 1) / 1][Stage 39:>
[Stage 40:>                                                         (0 + 1) / 1][Stage 41:>

[OK] All models trained

[OK] Expanded feature count = 846

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

25/10/08 02:15:07 WARN Instrumentation: [c11acd55] regParam is zero, which might cause numeri
[Stage 42:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:14 WARN Instrumentation: [9fe97b12] regParam is zero, which might cause numeri
[Stage 43:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:21 WARN Instrumentation: [3abdd974] regParam is zero, which might cause numeri
[Stage 44:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:27 WARN Instrumentation: [a123efba] regParam is zero, which might cause numeri
[Stage 45:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:34 WARN Instrumentation: [e59b3267] regParam is zero, which might cause numeri
[Stage 46:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:40 WARN Instrumentation: [d3b40fdb] regParam is zero, which might cause numeri
[Stage 47:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:46 WARN Instrumentation: [469a6821] regParam is zero, which might cause numeri
[Stage 48:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:53 WARN Instrumentation: [f20bbce1] regParam is zero, which might cause numeri
```

```
[Stage 49:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:15:59 WARN Instrumentation: [59f9a079] regParam is zero, which might cause numeri
[Stage 50:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:05 WARN Instrumentation: [6c66f0cf] regParam is zero, which might cause numeri
[Stage 51:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:11 WARN Instrumentation: [bfdd737f] regParam is zero, which might cause numeri
[Stage 52:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:18 WARN Instrumentation: [fd8c4979] regParam is zero, which might cause numeri
[Stage 53:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:24 WARN Instrumentation: [01b21cb8] regParam is zero, which might cause numeri
[Stage 54:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:30 WARN Instrumentation: [8b873b81] regParam is zero, which might cause numeri
[Stage 55:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:36 WARN Instrumentation: [6fba5c14] regParam is zero, which might cause numeri
[Stage 56:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:42 WARN Instrumentation: [28ea614a] regParam is zero, which might cause numeri
[Stage 57:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:49 WARN Instrumentation: [153d611e] regParam is zero, which might cause numeri
[Stage 58:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:16:55 WARN Instrumentation: [3a9f6a21] regParam is zero, which might cause numeri
[Stage 59:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:01 WARN Instrumentation: [2dc9d61f] regParam is zero, which might cause numeri
[Stage 60:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:07 WARN Instrumentation: [0144bf0b] regParam is zero, which might cause numeri
[Stage 61:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:13 WARN Instrumentation: [e4df611a] regParam is zero, which might cause numeri
[Stage 62:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:19 WARN Instrumentation: [10c887ba] regParam is zero, which might cause numeri
[Stage 63:>                                                         (0 + 1) / 1]
```

```
Newton solver.
25/10/08 02:17:25 WARN Instrumentation: [940163ed] regParam is zero, which might cause numeri
[Stage 64:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:31 WARN Instrumentation: [61e09548] regParam is zero, which might cause numeri
[Stage 65:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:37 WARN Instrumentation: [e4d30cd4] regParam is zero, which might cause numeri
[Stage 66:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:43 WARN Instrumentation: [1fd45cc8] regParam is zero, which might cause numeri
[Stage 67:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:50 WARN Instrumentation: [6a8a1b1b] regParam is zero, which might cause numeri
[Stage 68:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:17:56 WARN Instrumentation: [e7c0493f] regParam is zero, which might cause numeri
[Stage 69:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:02 WARN Instrumentation: [1e9e9b0c] regParam is zero, which might cause numeri
[Stage 70:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:09 WARN Instrumentation: [a8697996] regParam is zero, which might cause numeri
[Stage 71:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:16 WARN Instrumentation: [a1e7771d] regParam is zero, which might cause numeri
[Stage 72:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:22 WARN Instrumentation: [fd31bbdc] regParam is zero, which might cause numeri
[Stage 73:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:28 WARN Instrumentation: [46c12d0e] regParam is zero, which might cause numeri
[Stage 74:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:34 WARN Instrumentation: [4c0c6f28] regParam is zero, which might cause numeri
[Stage 75:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:40 WARN Instrumentation: [931bc367] regParam is zero, which might cause numeri
[Stage 76:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:46 WARN Instrumentation: [b8f663f8] regParam is zero, which might cause numeri
[Stage 77:>                                                         (0 + 1) / 1]
Newton solver.
```

```
25/10/08 02:18:52 WARN Instrumentation: [5bb0222a] regParam is zero, which might cause numeri
[Stage 78:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:18:58 WARN Instrumentation: [3627c5e6] regParam is zero, which might cause numeri
[Stage 79:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:19:05 WARN Instrumentation: [4c074494] regParam is zero, which might cause numeri
[Stage 80:>                                                         (0 + 1) / 1]
Newton solver.
25/10/08 02:19:10 WARN Instrumentation: [ac4c359d] regParam is zero, which might cause numeri
[Stage 81:>                                                         (0 + 1) / 1]

Saved: _output/glr_summary.csv

                                                                                25/10/08 02:1
Newton solver.

Saved: _output/polylr_summary.csv
```

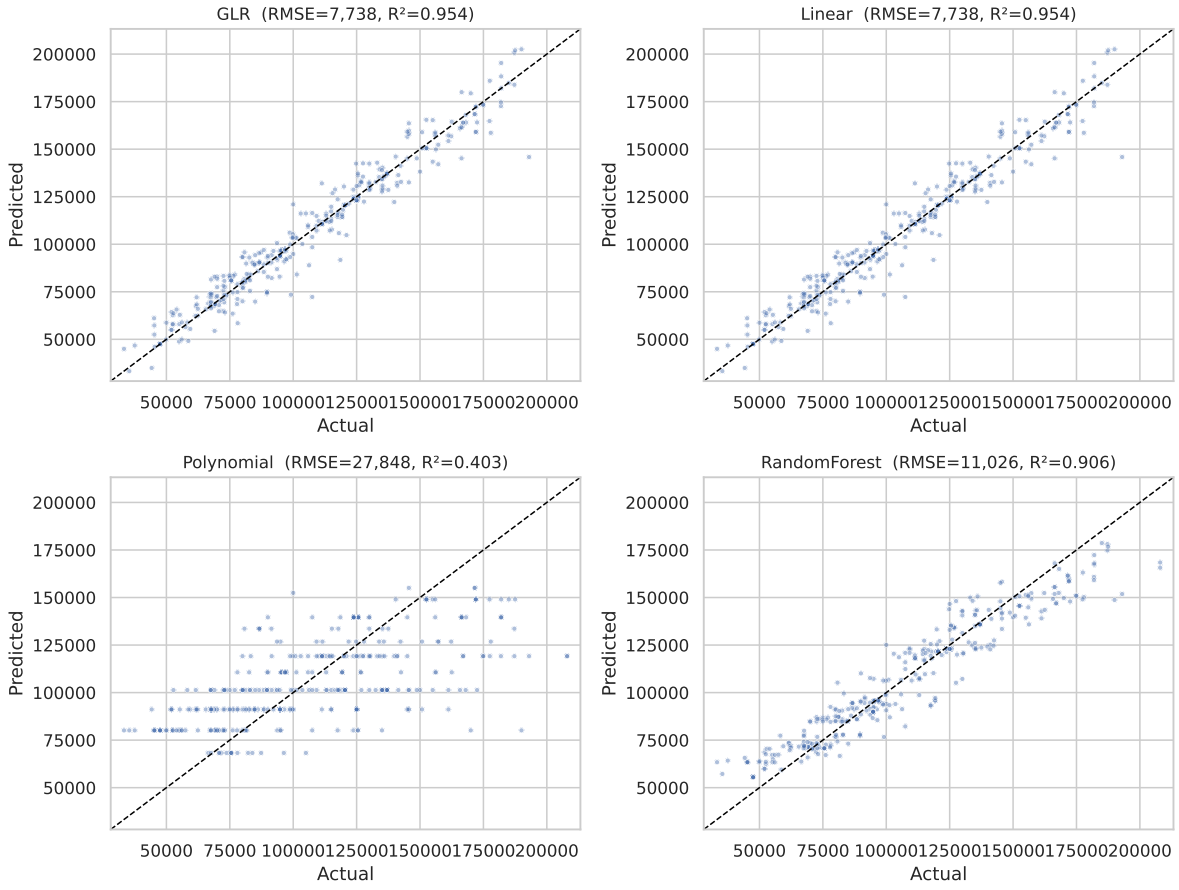
Interpretation of Polynomial Linear Regression. Adding a quadratic term in Min Yrs Exp does not improve generalization for this dataset. The linear term carries most of the signal, the smaller or non-significant t-value for the squared term is suggesting added variance with a limited predictive value.

```
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]

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

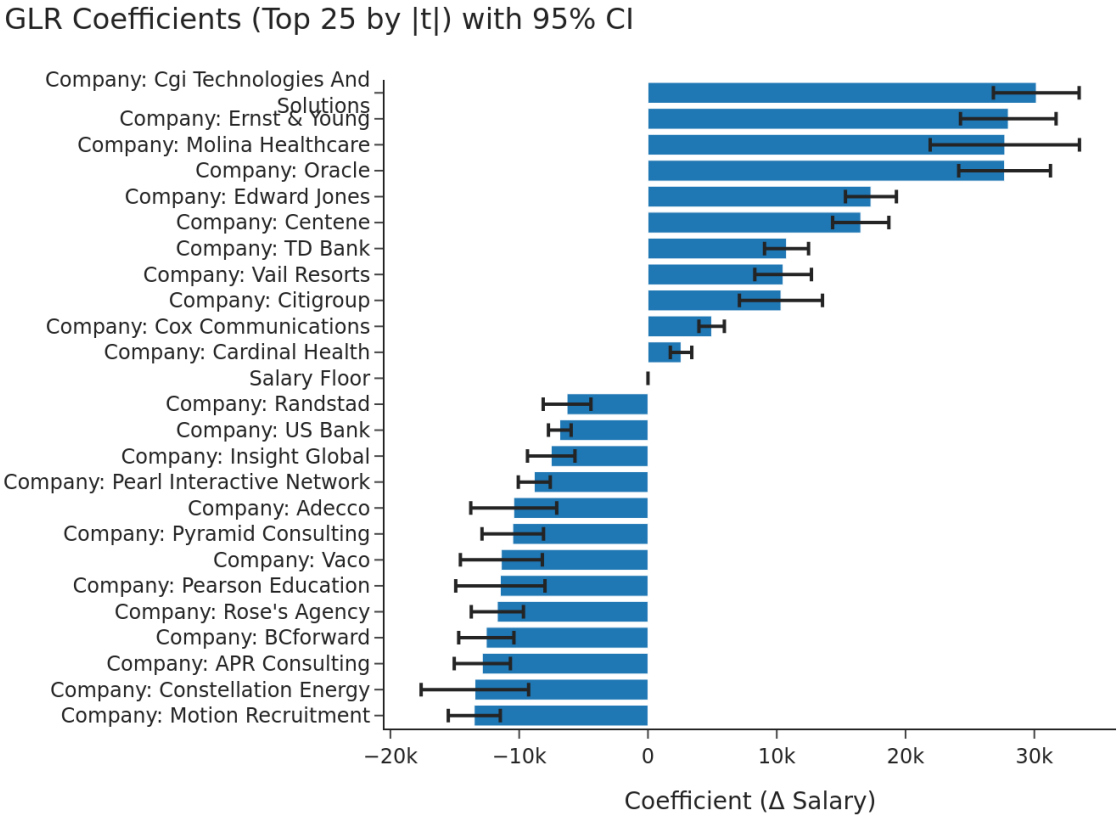


Saved: _output/actual_vs_pred_2x2.png

Interpretation of Model Comparison. All four models achieve similar predictive and performance accuracy. GLR and Random Forest show the lowest RMSE and comparable R^2. 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

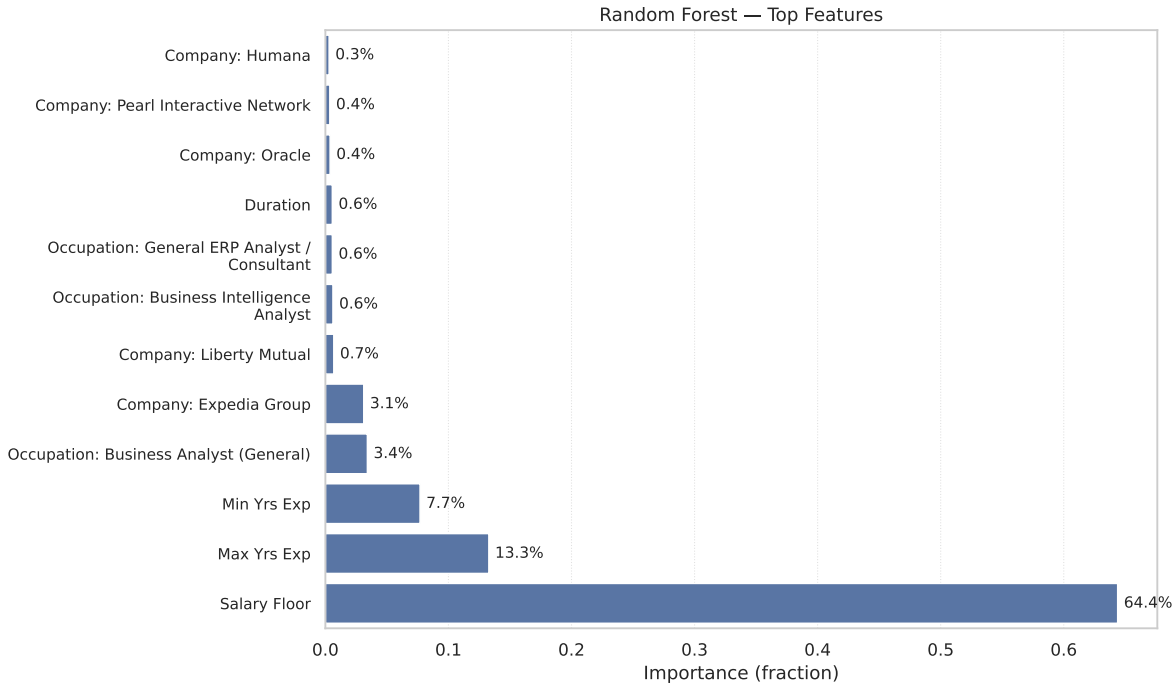
Saved: _output/glr_coefficients_ci.png



>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_13477/1242760667.py:74: UserWarning:

set_ticklabels() should only be used with a fixed number of ticks, i.e. after set_ticks() or
```



`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.