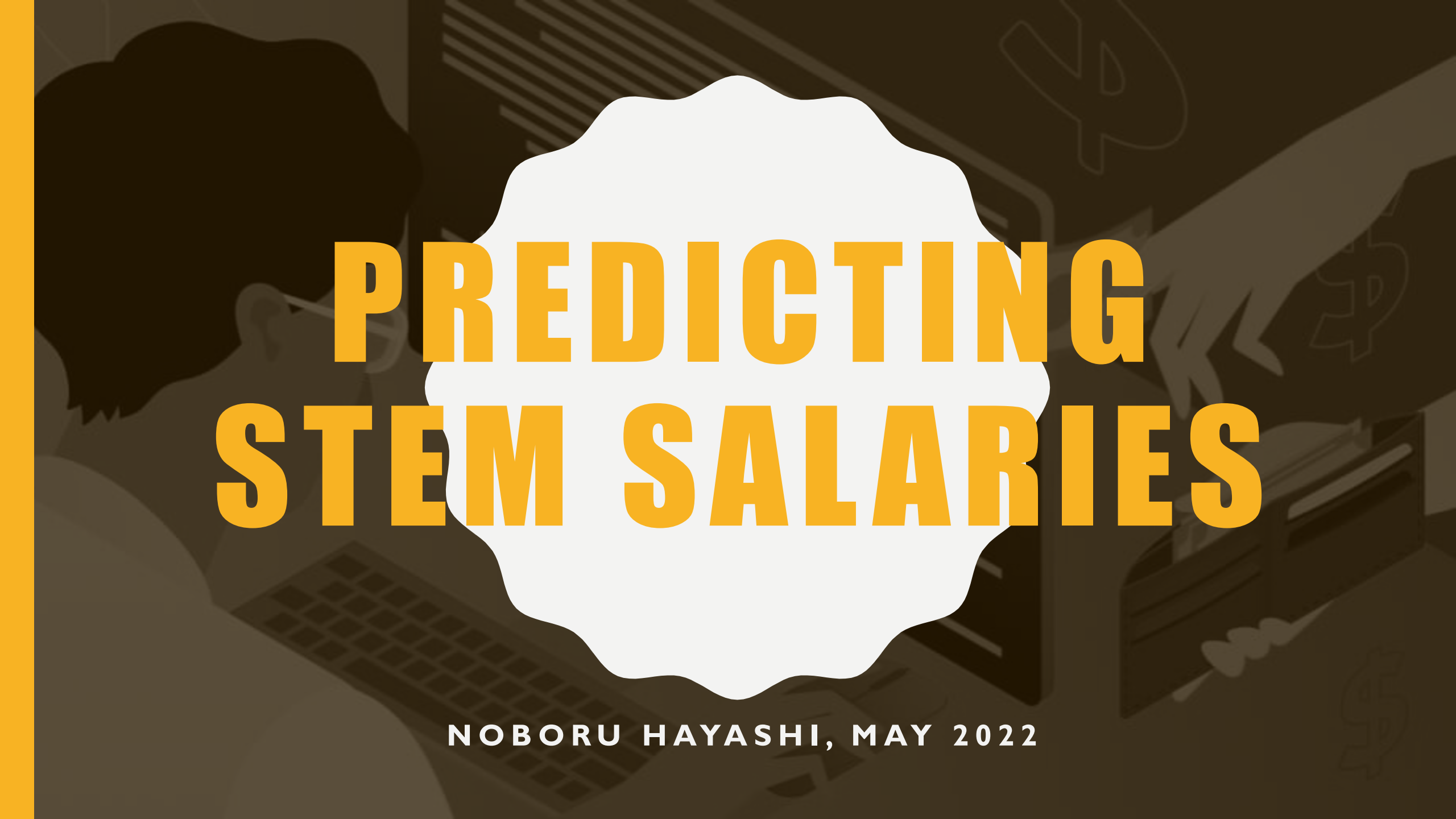
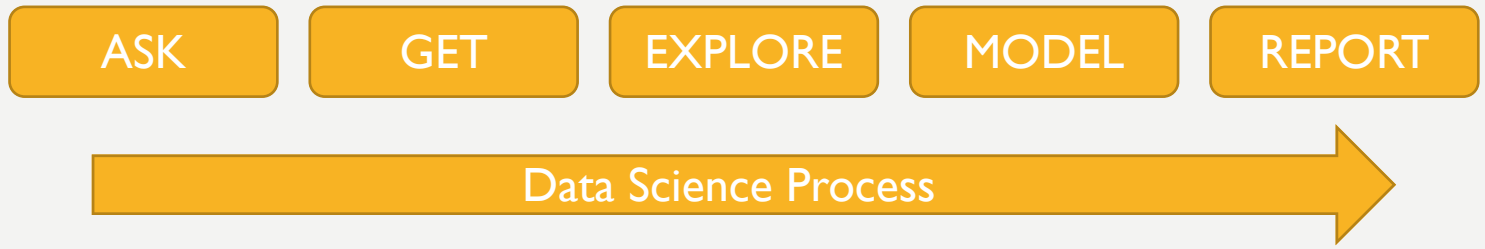


PREDICTING STEM SALARIES

NOBORU HAYASHI, MAY 2022

AGENDA

- 1. ASK – Motivation & Goal
- 2. GET - ETL
- 3. EXPLORE - EDA
- 4. MODEL – Model Building
- 5. REPORT – Result & Prediction



1. ASK – MOTIVATION & GOAL

- Personal curiosity about the salary of STEM workers from companies such as FAANG, as a Data Science major student
- Want to know more about the factors that possibly bring higher pay
- Knowledge of the market could help STEM specialists to understand his/her market value for making better career decisions
- Study Design: a linear regression model to predict salaries

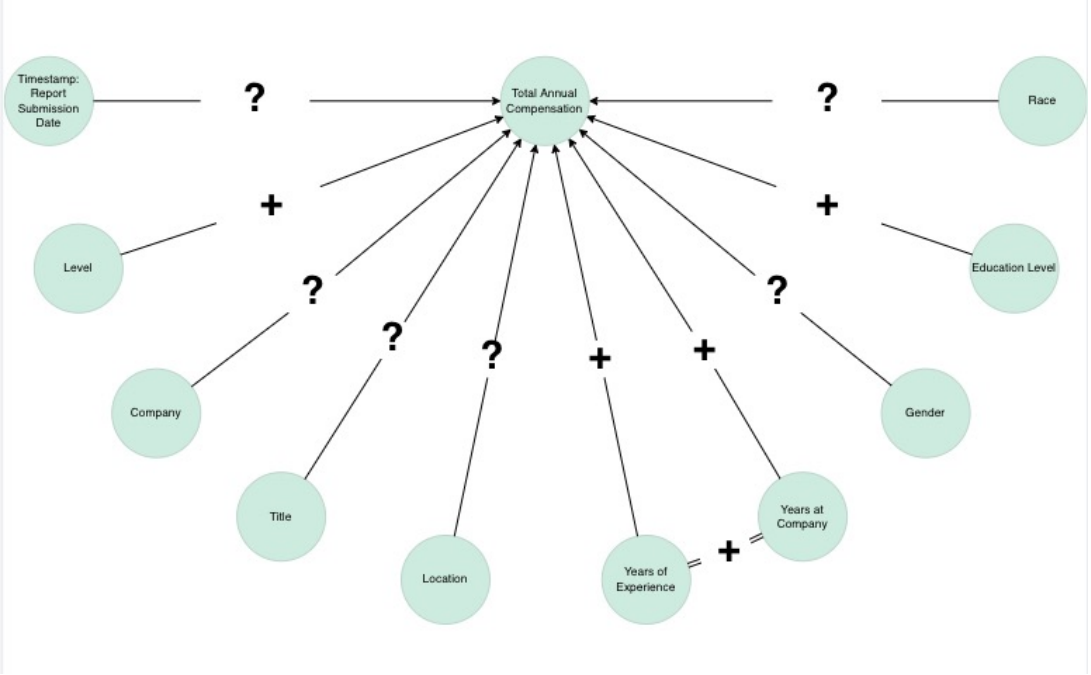
2. GET – ETL

- Data source: Data Science and STEM Salaries, Kaggle 2021
- Contains 62,000+ STEM salary records from scraped from levels.fyi, a website provides compensation package information of top companies
- The source file is in .csv format, can be systematically downloaded by using Kaggle API (Personal Kaggle credentials are required)
- The notebook file get_data.ipynb contains the codes to download the csv file.
- Once the csv file is download to local, it can be read as Pandas dataframe to facilitate analysis

```
od.download("https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries", "./data/")

Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username:

  nobobobo

Your Kaggle Key:

  ..............................

Downloading data-science-and-stem-salaries.zip to ./data/data-science-and-stem-salaries
100%|████████████████████████████████| 2.45M/2.45M [00:00<00:00, 15.2MB/s]
```

3. EXPLORE - EDA

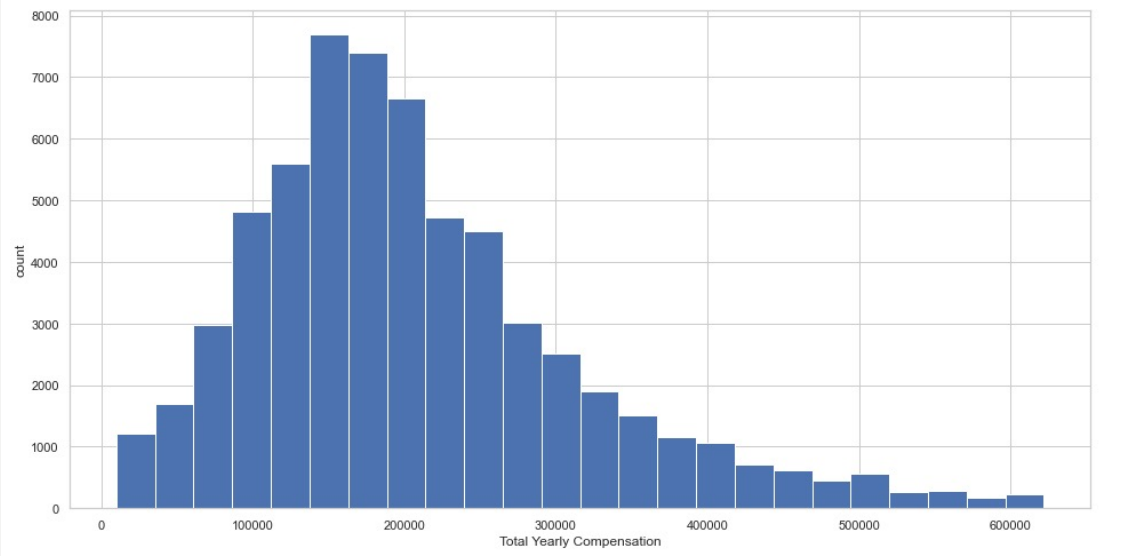
- Most features are categorical features
- Low-cardinality features, such as Gender, Race, Education level, are one-hot encoded. While high-cardinality ones are not
- Numerical features are Timestamp, Years of Experience, Years at Company
- Relationship examples:
  - Compensation vs. Years of Experience
  - Compensation from Common Companies vs. Uncommon Companies
  - Compensation in Common Locations vs. Uncommon Locations



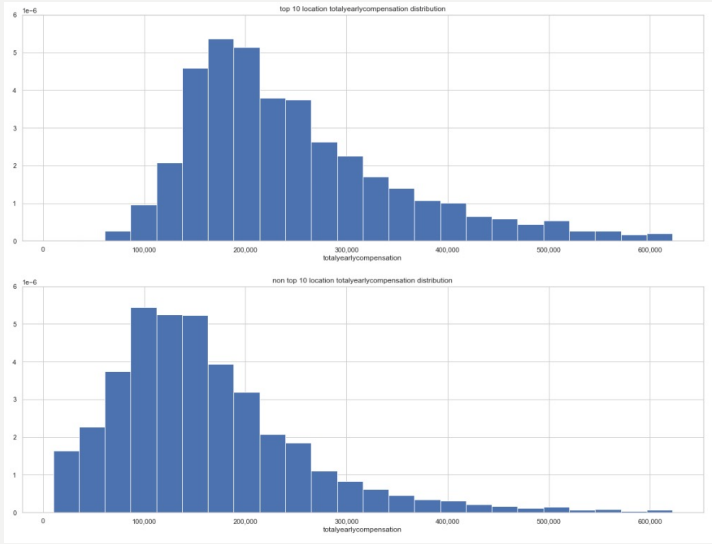
3. EXPLORE – EDA (CONT'D)

- Before EDA, the source file contains 62642 records with 29 columns
- Missing Values: No missing value in the target variable, Total Yearly Compensation. But 19540 records (31.19%) are missing gender value.
- Outliers: Removed about 1000 records (1.5%) with outlying total compensations
- High Cardinality String Fields: Some fields such as tag and other detail are text fields with high cardinality. These fields would require NLP to be used as regressors
- Since the target variable is numerical and most regressors are categorical, our pairwise comparisons mainly involved multiple histograms
- In EDA, created some grouping variables to group high cardinality features such as common companies vs. uncommon companies
  - The distributions of yearly compensations across these groups are both right skewed but the 'peaks' are slightly different
- Based on EDA, high cardinality fields require preprocessing before one-hot encoding, or other encoding method
- After EDA: 61718 rows, 60+ regressors

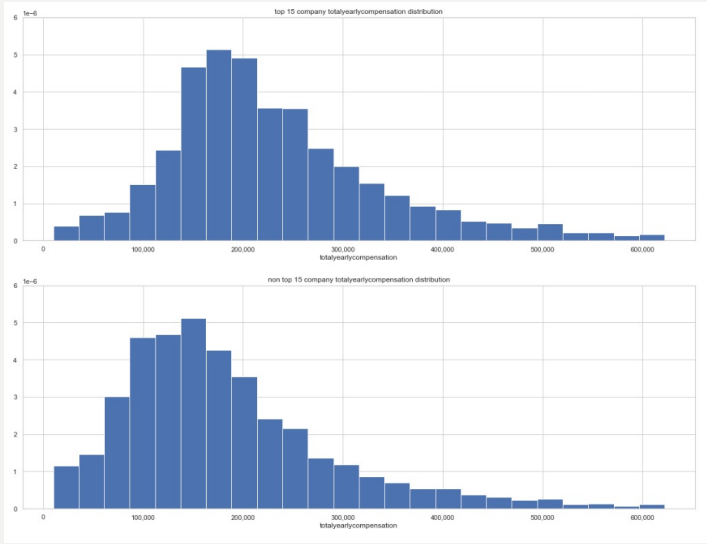
3. EXPLORE – EDA (CONT'D)



Histogram: Total Yearly Compensation



Total Year Compensation of 15 common companies vs. uncommon companies



Total Year Compensation of 10 common locations vs. uncommon locations

4. MODEL – MODEL BUILDING

- Mean Model (Null Model):
  - Mean(Total Yearly Compensation) = $207,197.61
  - 95% Error bounds ($\mu \pm 1.96 \cdot \sigma$) = ($ -6451.65, $ 420,846.88) => ($ 0.00, $ 420,846.88)

- Linear Regression Model:

$$\hat{y} = X\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where $\hat{y}$ is predicted total yearly compensation, $x_1, x_2, \ldots, x_n$ are regressors, and $\beta_0, \beta_1, \beta_2, \ldots \beta\_n$ are an intercept and corresponding coefficients

4. MODEL – MODEL BUILDING

- Initial model
  - Regressors:
    - mean encoded company & level
    - Timestamp converted as delta months from earliest month
    - years of experience, years at company
    - one hot encoded:
      - Job title, 10 common locations or else, 10 common countries or else, Race, Education
  - Score:
    - $R^2 = 0.827127$
    - $\sigma = \$\,45{,}341$
  - Strongest predictors:
    - Mean Encoded Company – Level: Same scale as target variable, its coefficient can be thought at % of importance
    - Job Title & Location (Countries): Flag fields for some job titles and countries having high coefficients
  - Possible transformations:
    - Log transform annual compensation and mean encoded company – level
    - Adding interaction term of years of experience and years at company

4. MODEL – MODEL BUILDING (CONT'D)

Coefficients		Mean	95% BCI Lo	95% BCI Hi
	β_0	-44295.781103	-46601.312683	-41988.975134
year_month_delta	β_1	389.443416	328.870839	439.864620
company_level_mean_enc	β_2	0.872630	0.866605	0.878654
yearsofexperience	β_3	1545.370088	1444.181238	1664.376586
yearsatcompany	β_4	-840.950580	-954.193836	-683.226859
Business_Analyst	β_5	-17465.438413	-19746.967283	-14765.310948
Data_Scientist	β_6	10798.881301	9187.791012	12690.226489
Hardware_Engineer	β_7	8664.095497	6969.266151	10626.804307
Human_Resources	β_8	-36436.205143	-41094.477084	-31476.346282
Management_Consultant	β_9	8237.513967	5716.231504	10497.366675
Marketing	β_{10}	-26455.519510	-29691.550451	-23588.032136
Mechanical_Engineer	β_{11}	-1062.404847	-3574.965343	2400.478178
Product_Designer	β_{12}	487.135724	-1991.391073	2446.032880
Product_Manager	β_{13}	3193.753464	2041.701028	5169.241205
Recruiter	β_{14}	-48040.676247	-52175.196173	-43766.037382
Sales	β_{15}	14589.712956	10094.353998	19995.917233
Software_Engineer	β_{16}	16657.003273	15750.357648	17497.093292
Software_Engineering_Manager	β_{17}	30044.938609	27392.900009	31887.891624
Solution_Architect	β_{18}	271.822970	-1905.153653	2338.878816
Technical_Program_Manager	β_{19}	-7780.394706	-10293.446438	-5478.897517
Australia	β_{34}	11179.461661	6954.397623	16205.826762
Canada	β_{35}	12690.194431	9536.192480	15731.603030
Germany	β_{36}	1351.197239	-2421.450809	4750.718780
India	β_{37}	-51374.595387	-55085.739650	-48693.470183
Ireland	β_{38}	-5800.441656	-11632.953911	-1989.877471
Israel	β_{39}	18894.847861	13773.889008	23980.454108
Singapore	β_{40}	14559.279729	8171.072088	20779.384047
Taiwan	β_{41}	-7492.382639	-14370.391778	-2777.443495
US	β_{42}	44597.011584	41797.997424	46725.614105
United_Kingdom	β_{43}	-1713.996936	-5809.214320	1757.400659
Race_Asian	β_{44}	353.584539	-1085.383177	1428.378096
Race_White	β_{45}	-2037.932535	-3603.154936	-534.834484
Race_Two_Or_More	β_{46}	-583.179384	-3634.032537	2461.328180
Race_Black	β_{47}	-4452.497672	-8438.275264	-2241.128294
Race_Hispanic	β_{48}	-8031.301280	-11108.262359	-5186.572242
Masters_Degree	β_{49}	3226.855756	1925.109018	4414.241659
Bachelors_Degree	β_{50}	190.682498	-798.851401	1436.459180
Doctorate_Degree	β_{51}	18666.114223	15942.772059	21508.273263
Highschool	β_{52}	1327.086893	-3531.982790	4699.909825
Some_College	β_{53}	-4105.572187	-9476.446554	81.491332

Metrics	Mean	Lo	Hi
σ	45341.553520	44930.299253	45765.766076
R^2	0.827127	0.823845	0.830776

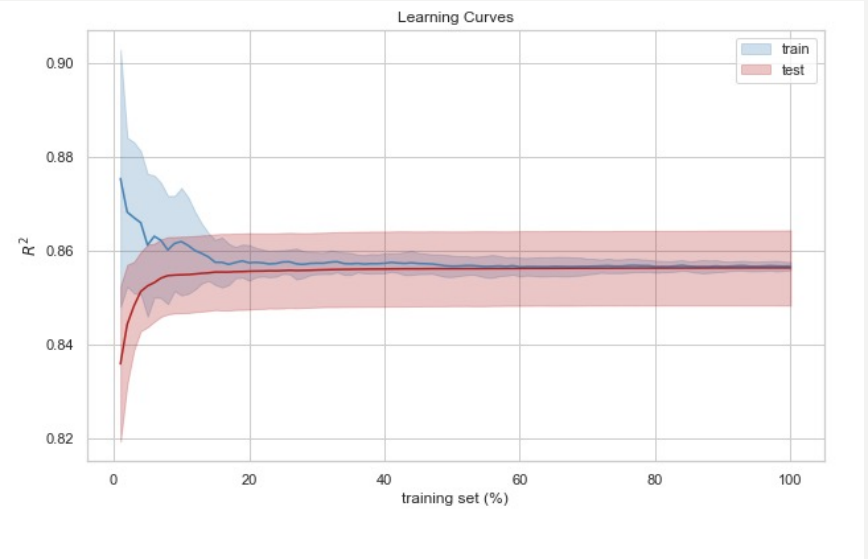
4. MODEL – MODEL BUILDING (CONT'D)

- Final Model
  - Target Variable:
    - Log(yearly compensation)
  - Regressors:
    - Log(mean encoded company & level)
    - Timestamp converted as delta months from earliest month
    - years of experience, years at company, and their interaction term
    - one hot encoded:
      - Job title, 10 common locations or else, 10 common countries or else, Race, Education
  - Score:
    - $R^2 = 0.856723$
    - $\sigma = 0.2233$

5. REPORT – RESULT & PREDICTION

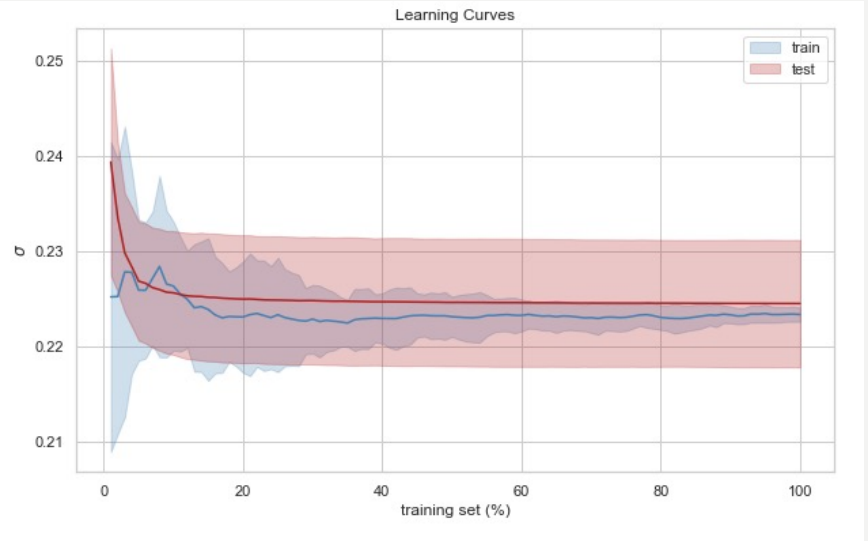
Cross Validation

- In order to understand better how our final model will perform, 3 rounds of 10-fold cross validation are performed.
- The means of R^2 and σ from 30 estimates are 0.86 and 0.22 which are quite aligned with the results of the final model.



Learning Curves

- Use increasing chunks of the training data to simulate getting more data for fitting the model, in order to examine the change in score as more data is used
- The scores converge relatively fast with lower amount of data
- Getting more data will not improve the model's performance



5. REPORT – RESULT & PREDICTION (CONT'D)

Prediction

Sample #1: a random person (data point) from the dataset

A Winston Salem, NC based White Male Human Resources person, working in Collins Aerospace as P6, graduated with MA degree, year of experience = 8, year at company = 8, the salary is submitted on April 17, 2021. His true total compensation is 164K.

Model Result:

Prediction: $ 130,933.23

95% Error bounds: $ 84,514.25 ~ $ 202,838.27.

5. REPORT – RESULT & PREDICTION (CONT'D)

Prediction

Sample #2: an 'imaginary' person with following info:

A Seattle based Asian Male Data Scientist, working in Amazon as L4, graduated with MA degree, year of experience = 1, year at company = 5, the salary is submitted on May 1st, 2022. His true total compensation for 2022 is 169K.

Model Result:

Prediction: $ 164,911.90

95% Error bounds: $ 106,449.10 ~ $ 255,482.97.


THANK YOU ☺