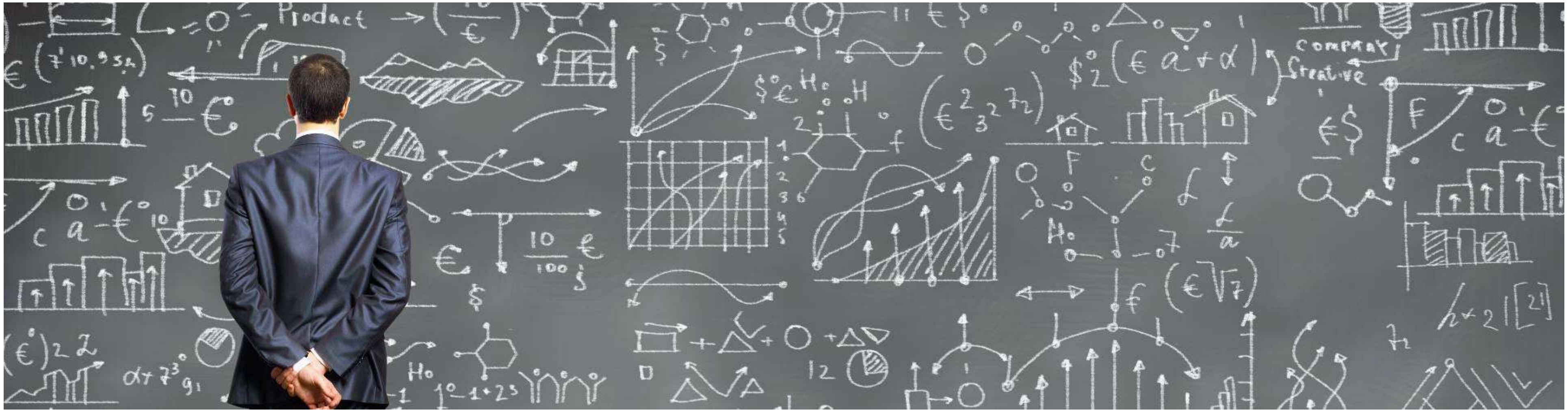
Gail Wittich



Take Home Data Science Exercise

Salary Predictions

13th July 2020

Agenda

Salary Predictions

Background / Business Problem

Predict Salaries with Accuracy

Predict the Salary for future job advertisements based on the salaries of previous job advertisements.

Situation	Many problems can result from incorrectly estimating the salary on offer when seeking new employees.
	• Attracting the right candidates.
	  • Offering too little will attract only under qualified candidates
	  • Offering too much may eliminate the advertisement from potential candidates search results.
	• Once employed:
	  • Underpaying staff can increase employee turnover
	  • Overpaying can trap employees when they are stale and ready to move up.
	The result being excessive time and money spent on recruitment and reduced employee productivity

Our professional reputation depends on good advice.

Complication	There are numerous factors that affect the salary for a given role, beyond the obvious Job Type: ▪ Experience required (in years) ▪ Industry ▪ Education level Location of the position Study Major Who the Employer is. These factors do not equally impact salaries, they may not have the same impact when combined with other factors. The objective is to develop a model that predicts salaries with an accuracy where the mean squared error is less than 320.

Executive Summary / Key Takeaways

Approach & Solution

- The relationship between salary and other factors was found to be:
  - JOB TYPE (the factor with the greatest impact on salary) CC 0.60
  - LEVEL OF EDUCATION (Degree) CC 0.40
  - YEARS OF EXPERIENCE and MAJOR are equally correlated CC 0.38
  - INDUSTRY and DISTANCE TO METRO are equally correlated CC 0.30
  - COMPANY had the least impact on salary (consequently eliminated) CC 0.0068
- Roles were grouped by 'jobType', 'degree', 'major', 'industry', generating new features.
- The mean of each group was found to have a greatest impact on the salary.

* CC = Correlation Coefficient

Including Years of experience in the groupings may have yielded even better results

Economic environment was deemed to be constant.

Data Set Characteristics

Salaries Training Data Set

Dataset Information

The Salaries Training Data Set:

- **train_features** 57.24 MB
  1 million records 8 Features
- Features and Labels:

  '**jobId**' 1 million unique values - This is the primary key

  companyId– 36 unique values (format: JOB999999999999)

  jobType – 8 categories, JANITOR, JUNIOR, SENIOR, MANAGER, VICE_PRESIDENT, CFO, CTO, CEO

  degree – 5 categories, NONE, HIGH_SCHOOL, BACHELORS, MASTERS, DOCTORAL

  major – 9 categories, NONE, LITERATURE, BIOLOGY, CHEMISTRY, PHYSICS, COMPSCI, MATH, BUSINESS, ENGINEERING

  industry – 7 categories, EDUCATION, SERVICE, AUTO, HEALTH, WEB, FINANCE, OIL

  yearsExperience – range 0 - 24

  milesFromMetropolis – range 0 - 99

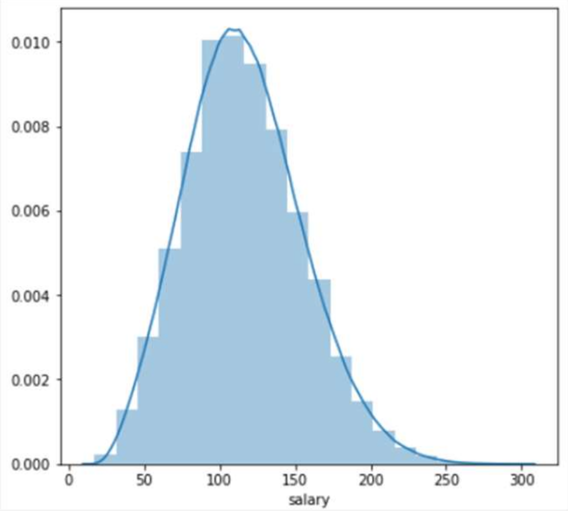
- **train_salaries** 3.94 MB
  1 million Records 2 Features
- Features

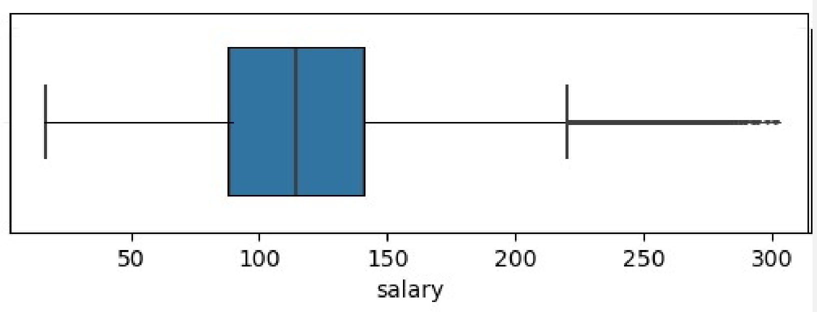
  jobId - 1 million unique values - This is the primary key

  salary - 279 unique values, range 0 – 24
  (expressed in $,000)

Dataset Visualizations

- '**jobId**' is the primary key.
- 5 records with invalid 'salary' data were removed.
- There were no duplicate records.

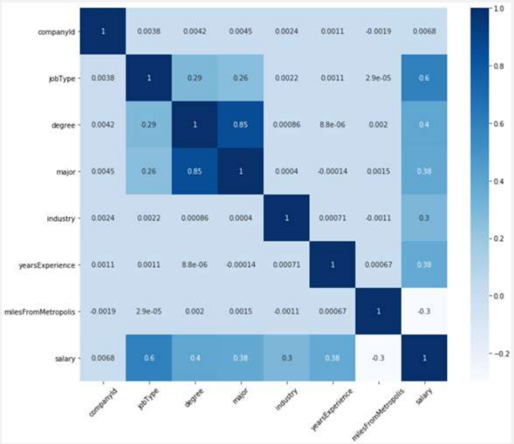




EDA – Exploratory Data Analysis

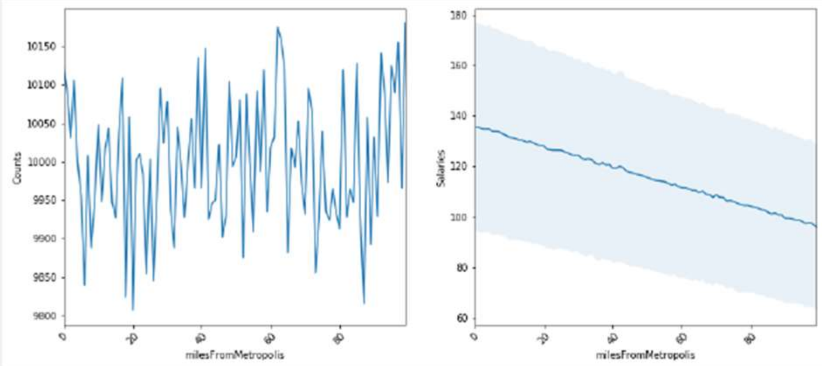
Correlation of Features to Target (Salary) (Numeric Features)

Correlation Map



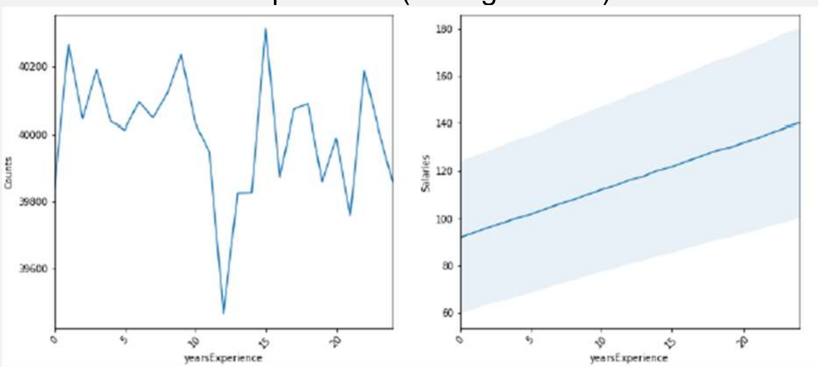
DISTANCE TO METRO

In general, salaries decrease with the distance to metropolis. (Weakest 0.3)



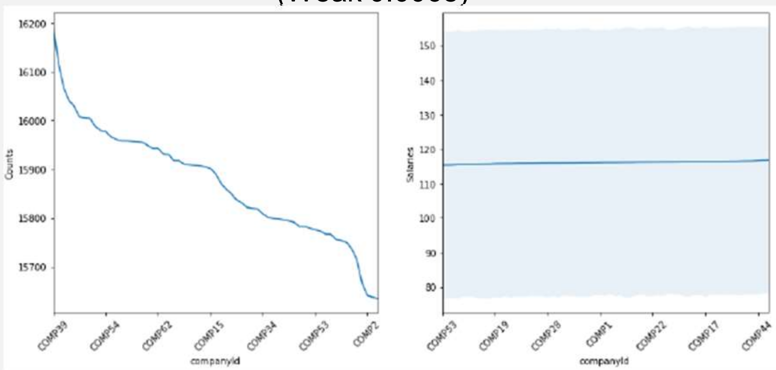
YEARS OF EXPERIENCE

In general, there is a clear correlation between salary and years of experience. (Strongest 0.38)



COMPANY ID

salary is weakly associated with companies. (Weak 0.0068)

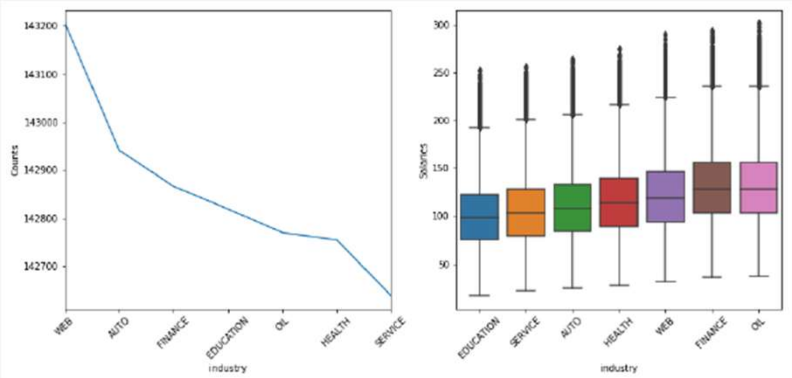


EDA – Exploratory Data Analysis

Correlation of Features to Target (Salary) (Categorical Features)

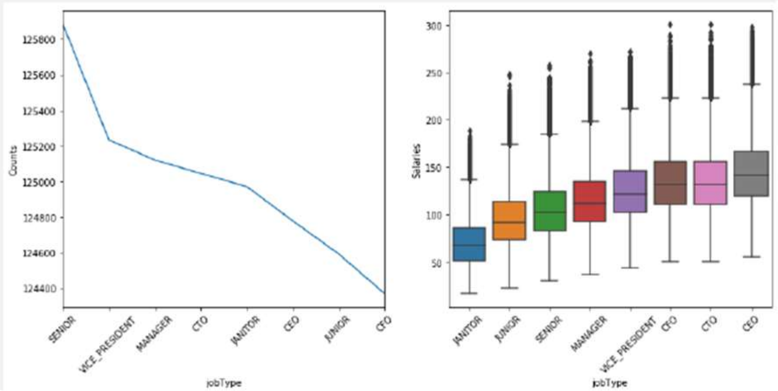
INDUSTRY

Oil, Finance and Web are generally better paid industries.
(Weakest)



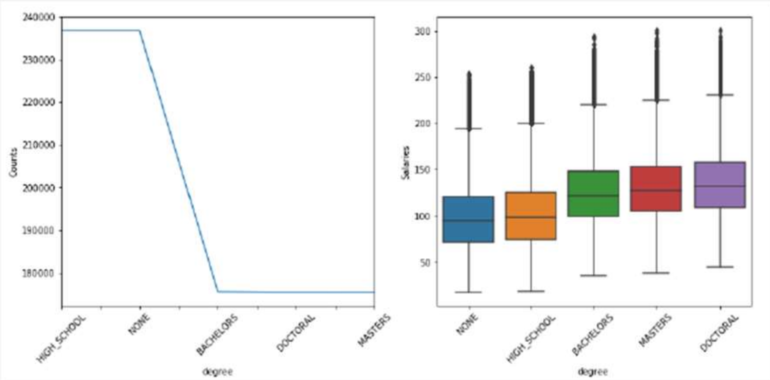
JOB TYPE

There is a clear positive correlation between job type and salary.
(Strongest)



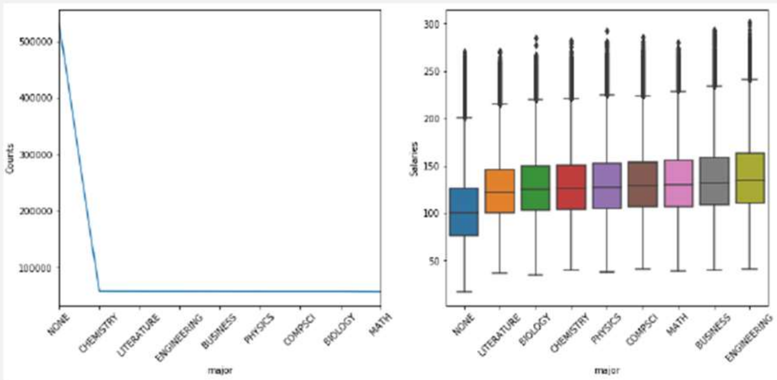
DEGREE

More advanced degrees tend to correspond to higher salaries.
(2nd Strongest)



MAJOR

People with majors of engineering, business and math generally have higher salaries. (3rd Strongest)



Data Cleansing & Pre-processing

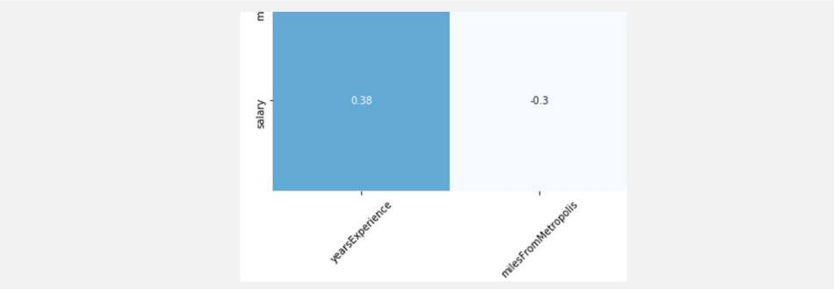
Quality Input , Quality Output

Categorical Features

	Correlation with Salary:
▪ companyId	0.0068 (EXCLUDED)
▪ jobType	0.60
▪ Degree	0.40
▪ Major	0.38
▪ Industry	0.30

Numerical Features

	Correlation with Salary:
▪ salary	(Target Feature)
▪ yearsExperience	0.38
▪ milesFromMetropolis	0.30



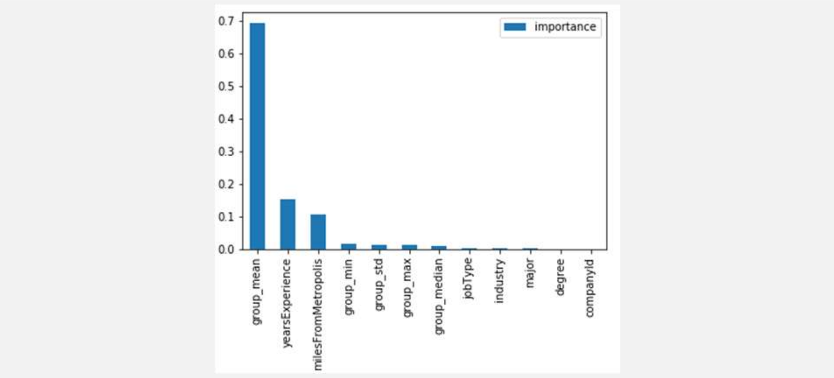
Feature Engineering / Dimension Reduction

Data was grouped according to:

- jobType   Degree
- Major   industry

The following **new statistical features** were calculated for the SALARY for each group:

- group_max   group_min
- group_std   group_median
- **group_mean (Feature Importance - 0.690169)**



Modelling, Tuning & Evaluation

Accurately Predict Salaries when Adverting Vacancies

Model Selection

- **Supervised Machine Learning** algorithms, specifically
- **Regression** and **Ensembles** of Regression Algorithms suit our data and goal.
- 3 models were selected:
  - LinearRegression — Sometimes simple is best
  - RandomForestRegressor — Offers Improved accuracy and control over-fittings
  - GradientBoostingRegressor — Can optimise on Least squares regression.

Hyper parameter tuning

- RandomForestRegressor – 60 estimators, max depth of 15, min. samples split of 80, max. features of 8
- GradientBoostingRegressor – 40 estimators, max. depth of 7, loss function used was least squares regression.

Cross validation

- 5-fold cross validation, scoring = neg_mean_squared_error

Model Evaluation

Models were evaluated using **Mean Squared Error (MSE)**
The lower the MSE the better the prediction.

Our goal is to achieve an MSE of less than 320

Benchmark Model – LinearRegression	MSE: ~399.8
LinearRegression (after Feature Engineering)	MSE: ~358.2
RandomForestRegressor	MSE: ~313.6
GradientBoostingRegressor	MSE: ~313.1

Model Performance Results

GradientBoostingRegressor model was selected.

40 estimators, max. depth of 7, loss function used was least squares regression.

Achieving a **22% improvement** over the baseline model.

.

Analysis Results & Recommendations

Where We Are

[Define/describe here the key results and recommendations of the Analysis
- List and strength of the key predictors
- Performance of the model
- Business recommendation and outcomes]

Result #1

- Mean Salary when grouped by Degree, Major, Job Type and Industry became a grater indicator of Salary than another feature.

Result #2

Result #3

Next Steps & Improvements

Good can get Better

- Expand Group Statistics criteria

Project/Approach Improvements

1. Given the strength of the relationships between Salary and the following features:

   - YEARS OF EXPERIENCE

   - DISTANCE TO METRO

   future models should include summary statistics for groups, that include theses features.

2. The high correlation between degree and major will likely have caused collinearity problems in the prediction in this model. Future models should exclude either DEGREE or MAJOR.

Lessons learned

1. Greater insights can be gained by feature generation than the data provides on its own.

2. Collinearity must be considered during Feature Selection.

Appendix

Assumptions

Analysis

1. Employment conditions, other than Salary, are not taken into consideration and are therefore assumed to be the same and not affect the Salary offered.

Results

2. It is assumed that the time period between job advertisements is not so great that the economic environment and therefore salaries will have changed.

Data Science Approach

1. Understand the problem	▪ Never forget which business problem you are trying to solve and the business objectives.
2. Explore the data	▪ Exploratory data analysis to understand the quality of the data (i.e. missing fields), the shape of the data (size, number of features, type of features), the statistic profile of the data (i.e. outliers, distribution etc.)
3. Cleanse the data	▪ Clean any data quality issues: garbage in, garbage out
4. Preprocess the data	▪ Transform the data or engineer new features, if necessary, to gain more insights
5. Metrics and Modeling	▪ Model creation, evaluation and selection
6. Evaluate findings	▪ Are they logical and do they make sense? Is the modeling approach used appropriate?
7. Iterate and Refine	▪ Refine analysis and fine tune models and findings
8. Communicate clearly	▪ Simple and straightforward messaging linking the results to the business outcome. ▪ Assumptions stated.

Code is clean, easy to read and the analysis is repeatable

Development Environment

https://github.com/gwittich/SalaryPredictionProject
- Python
- Libraries: Scikit-Learn, Pandas, Numphy.
- Other

Code

- Windows, Ubuntu
- Anaconda

Environment

- Seaborn,
- MatPlotLib
- SciPy

Visualization

- .CSV
- Pickle

Data