Executive Summary

Background

- CF (Contracting Firm) provides technology and scientific solutions
- Skilled people are a key component of the solutions
- CF would like to win more contracts by being more competitive in the hiring market

Project Objectives

- Determine the industry factors that are the most important in predicting salaries
- Factors that distinguish job categories and titles from each other. Can required skills accurately predict job titles?

Scope

Data related job postings including data scientist, data analyst, research scientists, business intelligence within a geography

Outcomes

The project outcomes are two models of differing scores for accuracy and recall.

Features that indicate above or below median salaries

Feature	Above Median	Below Median
Job Title	Software Engineer Engineer Machine Learning Statistician Data Scientist	Associate Research Engineering xxxx Analyst
Job Summary	Learning Data Scientist Work Knowledge Experience Data Analysis Ability Quality	Analyze xxx Research Develop Design
Location	az pa washington dc	coral xxx fl austin xxx
City	phoenix philadelphia washington dc	portland miami houston
Company Name	department xxx cancer xxx	university xxx state

Model 1 – Logistic Regression

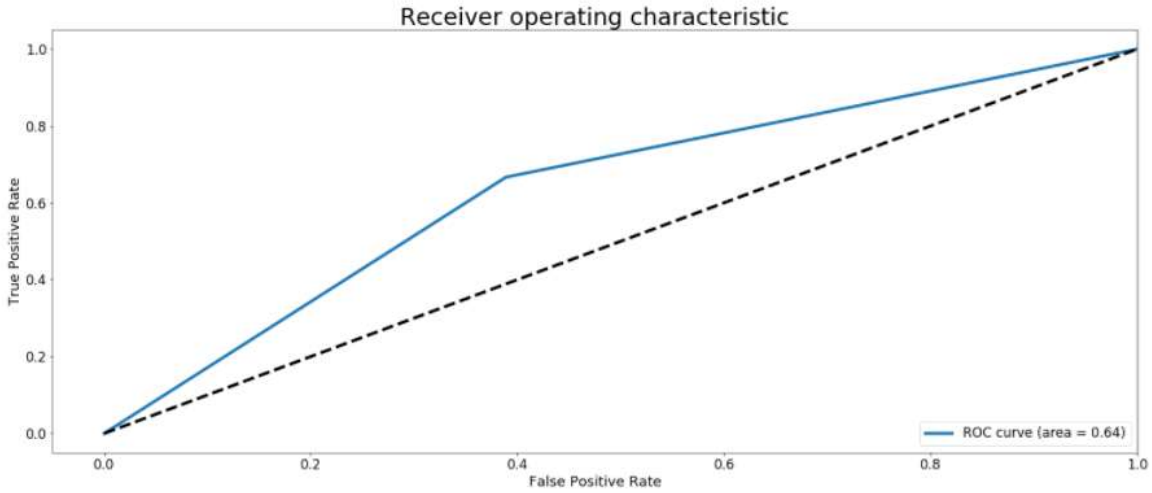
```
Score: 0.6388888888888888

              Predicted_below_median  Predicted_above_median
Below_median                      11                       7
Above_Median                       6                      12

              precision    recall  f1-score   support

below_median       0.65      0.61      0.63        18
above_median       0.63      0.67      0.65        18

 avg / total       0.64      0.64      0.64        36
```



Model 1 Commentary

Score

Model scored below the test set indicating need for further tuning or more data

Confusion Matrix Row = Ground Truth; Column = Prediction

- 11 is below median real and predicted correct
- 7 is below median (real) and predicted above median (incorrect)
- 12 is above median real and predicted correct
- 6 is above media real and predicted below median (incorrect)

Classification Report

- Precision = how useful is the model; how many positives are really true (TP/(TP+FP)) or TP/Predicted_True
- Recall/Sensitivity = how complete is the model; of all actual, % correct (TP/(TP+FN)) or TP/Real_True
- f1 = weighted harmonic mean of precision and recall. best = 1, worst = 0
- support = # of observations for each class

Model 2 – Gradient Boosting

```
Score: 0.8333333333333334
```

	Predicted_below_median	Predicted_above_median
Below_median	17	1
Above_Median	5	13

	precision	recall	f1-score	support
below_median	0.77	0.94	0.85	18
above_median	0.93	0.72	0.81	18
avg / total	0.85	0.83	0.83	36

Model 2 Commentary

Using Gradient Boosting with no parameter tuning

Score

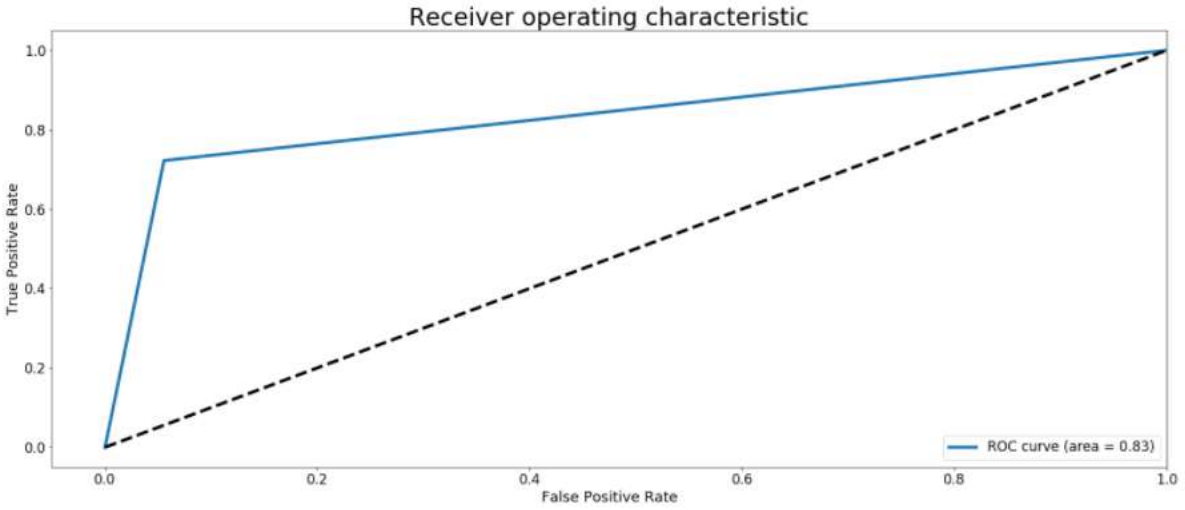
Better than logistic regression

Confusion Matrix Row = Ground Truth; Column = Prediction

- 17 is below median (real) and predicted correct
- FP: 1 is below median (real) and predicted above median (incorrect)
- 13 is above median (real) and predicted correct
- FN: 5 is above median (real) and predicted below median (incorrect)

Classification Report

- Precision (accuracy) - Model predicts above median more accurately than below median
- Recall (completeness) - Model is more complete for below_median
- f1 score is balanced between both above and below median



Limitations and Risks

- While the scraping provided over 3,382 job postings, after de-duplication, job posting with data was less than 200.
- Only job postings with salary was used. Potentially this may provide a sample bias for location, company name and job title.
- Data was taken from the US and extrapolation to a new country eg Singapore may not be valid

BONUS PROBLEM

Your boss would rather tell a client incorrectly that they would get a lower salary job than tell a client incorrectly that they would get a high salary job. Adjust one of your models to ease his mind, and explain what it is doing and any trade-offs. Plot the ROC curve.

Requirement

- Achieve Zero False Positive(high specificity) - Below Median (real) and Zero predicted Above Median
- If it is a False Negative - Above Median (real) and predicted below median, that is acceptable
- Trade off is that a higher specificity means lower sensitivity (model is less accurate)

Approach

1. Optimise for sensitivity using GridSearchCV with the scoring argument
2. Adjust the decision treshold to identify the operating point

precision_recall_curve and roc_curve can be used to visualise the sensitivity-specificity trade off in the classifier. Helps inform where to set the decision threshold to maximize sensitivity or specificity. This is called the 'operation point' of the model.

.predict_proba() and .decision_function() returns the raw probability that a sample is predicted to be in a class. This is an important distinction from the absolute class predictions returned by calling the .predict method.

Approach

Extraction

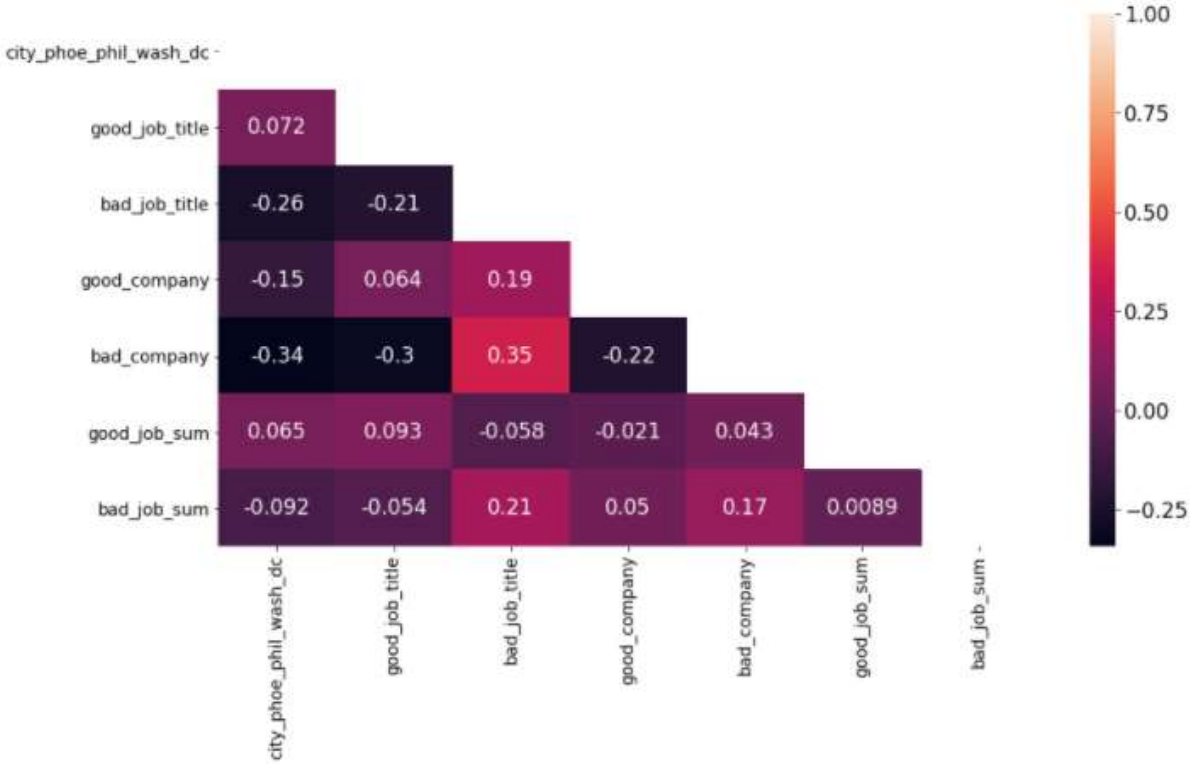
Indeed.com job site in the US (3,382 job postings, 6 features)

Field	Description	Example
Job Title	Data Scientist	Data Scientist - Journey Analytics
Company Name	Company Name	DiMeo Schneider & Associates
Location	Company Location	New York NY 10022 (Midtown area)
Salary	Salary	$90,000 - $115,000 a year
Job Summary	Job Summary	We are looking for a creative and innovative data scientist to drive strategic initiatives that will propel the future growth of Groupon...
City	City	New York

EDA Process

1. Drop duplicate rows
2. As salary data is right skewed indicating high outliers, convert Salary range data to median salary and create new salary column 'Median Salary'
3. Create new salary column - 'Above Median Salary' . Above Median Salary = 1, Below Median Salary = 0
4. Explore job characteristics with NLP
    a. Use n-gram range 1 - 3 words
    b. Words appear in at least 5% of postings
    c. Ignore stop words (stop_words = 'english')
5. Identify n-grams for above_median and below_median jobs
6. Use Ratios for the relative appears for each n-gram in above_median and below_median
7. Do this for 'job_title', 'summary', 'location', 'city','company_name'
8. Feature Engineering
    a. city_phoe_phil_wash_dc
    b. good_job_title
    c. bad_job_title
    d. good_company
    e. bad_company
    f. good_job_sum
    g. bad_job_sum

9. Check the correlation between the features above for separation of the variables



Model

- y = above_median_salary or below_median_salary
- X = features from #8 above
- Determine baseline = 0.5

Model 1 - Logistic Regression

a. Model Score 0.75
b. Predict Score 0.64
c. Precision 0.64
d. Recall 0.64
e. f1 score 0.64
f. Confusion Matrix
    - 11 is below median real and predicted correct
    - 7 is below median (real) and predicted above median (incorrect)
    - 12 is above median real and predicted correct
    - 6 is above media real and predicted below median (incorrect)

Model 2 - Gradient Boosting Classification

a. Model score 0.83
b. Predict score 0.83
c. Precision 0.85
d. Recall 0.83
e. f1 score 0.83
f. Confusion Matrix
    - 17 is below median (real) and predicted correct
    - FP: 1 is below median (real) and predicted above median (incorrect)

- 13 is above median (real) and predicted correct
- FN: 5 is above median (real) and predicted below median (incorrect)