1. **Problem statement and hypothesis**

I am seeking to predict a job’s salary based on its job advertisement. This is a Kaggle Competition that is closed that was proposed by a company called Adzuna.

Creating this prediction model would make improvements in the experience of users searching for jobs and help employers and jobseekers figure out the market worth of different positions.

1. **Description of your data set and how it was obtained**

The data for [Job salary prediction](http://www.kaggle.com/c/job-salary-prediction) consists of several columns, which are mostly either text or categorical.

I downloaded the train and test csv files here: <https://github.com/ajschumacher/gadsdata/tree/master/salary>.

"Train" is the data I explored and used to model, and had the response value "SalaryNormalized". "Test" has all of the same fields as "train" except the response value (salaries) , and my goal is to predict that response.

Below is a table of the data column names and descriptions.

|  |  |
| --- | --- |
| ID | A unique identifier for each job ad |
| Title | A freetext field supplied to us by the job advertiser as the Title of the job ad.  Normally this is a summary of the job title or role. |
| FullDescription - | The full text of the job ad as provided by the job advertiser.  Where you see \*\*\*s, we have stripped values from the description in order to ensure that no salary information appears within the descriptions.  There may be some collateral damage here where we have also removed other numerics. |
| LocationRaw | The freetext location as provided by the job advertiser. |
| LocationNormalized | Adzuna's normalised location from within our own location tree, interpreted by us based on the raw location.  Our normaliser is not perfect! |
| ContractType | Full\_time or part\_time, interpreted by Adzuna from or a specific additional field we received from the description advertiser. |
| ContractTime | Permanent or contract, interpreted by Adzuna from description or a specific additional field we received from the advertiser. |
| Company | The name of the employer as supplied to us by the job advertiser. |
| Category | Which of 30 standard job categories this ad fits into, inferred in a very messy way based on the source the ad came from.  We know there is a lot of noise and error in this field. |
| SalaryRaw | The freetext salary field we received in the job advert from the advertiser. |
| SalaryNormalised | The annualised salary interpreted by Adzuna from the raw salary.  Note that this is always a single value based on the midpoint of any range found in the raw salary.  This is the value we are trying to predict. |
| SourceName | The name of the website or advertiser from whom we received the job advert. |

1. **Description of any pre-processing steps you took**

I created a dummy variable (an artificial variable created to represent an attribute with two or more distinct categories/levels) for Contract Type and Contract Time.

I also used “train.ContractTime.fillna('was\_na', inplace=True)” to replace nulls with new values.

1. **What you learned from exploring the data, including visualizations**

When I square rooted the Salary Normalized data, I got a standard bell curve.

1. **How you chose which features to use in your analysis**

To decide which features to use in my analysis, I wanted to see which x values had relationships to my y value, salary. I also wanted to see how strong that relationship was.

Because Contract Type, Contract Time, Location and Category are categorical, I could not use a scatter plot.

I would like to use some of the ad text as features. I used count vectorizer and

I would like to use KNN to calculate similarity between titles, average location and company salaries.

I would also like to see how I can do using less information. An advantage of using less information is that you cannot overfit the data as much. I would like to leave out some fields one at a time and find the mean errors without Title, Company, Category, etc and see how low a MSE I can get.

1. **Details of your modeling process, including how you selected your models and validated them.**

I am using a multiple linear regression model.

I hope to use a Random Forest model as well, as soon as I know what other models I can use.

1. **Your challenges and successes**

TBD

1. **Possible extensions or business applications of your project**

TBD

1. **Conclusions and key learnings**

TBD