**Literature Report**

By Hritika Rahul Mehta

SRN-PES220180024

Title- Salary Prediction in the IT Job Market with  
Few High-Dimensional Samples: A Spanish Case Study

Ignacio Martın , AndreaMariello , RobertoBattiti , Jose ́ Alberto Hernandez

**Summary**

This paper focuses on predicting the salaries of jobs offered on an online portal called “Tecnoempleo” that specialises in IT jobs for people residing in Spain. E-recruitment of candidates is a hot topic of research and there are many different approaches like automatic evaluation of CV, ranking of the skills of candidates which are used by companies to aid in the choice of a candidate.

The features collected by web scraping are too vast so a manual data pre processing approach is used to improve prediction accuracy. Salary prediction by ranges is the main focus of this paper, several models like – linear models, logistic regression ,K-nearest neighbours, multi-layer perceptrons ,support vector machines, random forest, adaptive boosting with decision trees along with ensembles of the models stated before are considered in order to formulate a solution. Data inspection phase of this paper includes creation of distribution plots(bar charts) of number of companies against the number of job posts, most frequent job positions and geographical distribution of jobs posted. Boxplots are also used to showcase salaries as a function of years of experience and salaries as a function of education. Various models as mentioned previously are then compared based on R2 value,F1-score among other parameters, the voting classifier that includes the 3 best performing models is shown to perform the best for the given classification problem.

Domain and geographical restrictions of the considered website make the prediction task challenging along with the volume of features collected over a period of 5 months ,including missing data and fewer job posts. A voting classifier is fitted on data which is manually pre-processed and normalized between a scale of 0-1;the parameters for which are obtained by using grid search.

**Assumptions of the paper**

In this work data is collected from an e-Recruitment website with geographical limits to Spain;the posts for jobs not based in Spain were removed.It contained a large number of job posts but only a few of those contained the offered salary which includes close to 4000 posts. Here the missing values were removed because in most cases there was no default value with which they could be substituted. Information which is unstructured like the title of the job post, description of offered post and requested profile were also removed. Different keywords related to the software versions required by various companies were removed as well. Keywords which appear less than 10 times are considered noise and thus eligible for removal from the dataset. These comprise the assumptions of the paper considered for this report.

**Solution Approach**

Salary classification according to range results in a better categorisation of job posts which therefore helps end users to navigate the website easily is the main focus of this paper. The phases of solution include

* **Data collection**- A Python based web crawler is used to gather information from the “Tecnoempleo” website for 5 months between December 2015 and April 2016.
* **Data cleaning**- The job postings with lot of missing features are removed from the dataset along with others that do not have Spain as the geographical location of the job.
* **Manual feature engineering**- Some irrelevant features are discarded and others are modified as follows – the *dedication* feature is transformed into maximum number of week hours; *incentives* is transformed into a binary variables; the approximate number of employees are retrieved for each of the 488 companies that have posted a job offer, this information is easily obtained from the public profiles of companies and for the ones that it is not available it is filled with the median value of workers from the other companies; textual representation of the working experience is converted to an average number of years; a new variable of per capita gross product in the geographical region of the each company is introduced. Most of the keywords are manually entered into the website so in order to maintain data integrity they unify semantically equivalent words. Finally the variables are standardised to values between 0 and 1 using their range of value.
* **Dataset Inspection***-* This phase is similar to data visualisation and various plots based on salary are plotted and few preliminary models are used to provide a description of data these include K-means clustering, linear regression and regularised linear regression which is a Ridge regression model in this case. It can be inferred that salary expectations increase with the number of years of experience and there is no particular trend concerning education but it is clear that experience is more rewarded in the IT industry. Programmer positions make up more than half of the available posts.
* **Automatic feature selection-** The filter method of X-MIFS which is a method of feature selection based on maximisation of mutual information between features and output variable(salary range).This method is chosen due to it’s scalability to high dimensional data and independence from classifiers which are used. Automatic feature selection helps to test for the existence of an optimal feature subset that is difficult to obtain from manual methods.
* **Model Selection –**A preliminary grid search on 90% of data by using 3-fold cross validation is used to find the optimal hyper parameters of various machine learning models. Generalised linear models are compared with support vector machines with a linear kernel. For KNN models, algorithms based on Manhattan l1 norm distance and Euclidean l2 norm distance are compared. Stochastic gradient descent is used for weight optimization of Multi Layer Perceptron model .Non-linear SVM models are evaluated using radial basis function along with sigmoid kernels .Random Forest classifiers and Adaptive Boosting classifiers use Gini impurity criterion or Information Gain criterion. Vote and Vote3 models soft voting is used.
* **Model training and validation-** The models stated above are trained and cross validated to find the best classifier among them.
* **Model comparison –** Different scores used to compare algorithms are:
* **Accuracy -** The percentage of correctly classified samples.
* **Precision –** Itis the fraction of true positives among true positives and false positives.
* **F1 score-** This is the harmonic mean of precision and recall.
* **Area under precision recall curve**
* **Area under ROC curve –** it shows true positive rate versus false positive rate.

Based on the above criteria Vote-3 is the most robust model with highest precision and accuracy. Not only that it has a high F1 score and continues to perform well when the context is changed to binary classification.

**Relevance to Project**

Our project is based on a dataset scraped from Glassdoor which like the website considered in the paper is a e-Recruitment website with geographical limitation to the United States. We need to employ similar feature engineering to our dataset concerning the Salary estimate and Size columns by converting them into a numeric form. It is very clearly concluded about how salary classification into a range makes job profiling easier and hence a website easier to use for a person hence we could also attempt a classification approach to our problem. Few columns with missing data could be dropped as well due to less impact on the outcome variable as followed in the above approach.