Job Salary Prediction

a theoretical study to predict job salaries based on job descriptions

Group Members

SADMAN SHAWMIK
Section 03 — 15301036

LAMIA TASNIM
Section 02 — 13301121

SHINY RAISA
Section 03 — 15301047

Introduction

- Since the job market is very competitive, we want to build a prediction engine that can spit out salaries based on different criteria like company, location, position, etc. The prediction model will be able to:

  — Greatly improve job seekers' search experience

  — Assist employers and job seekers in estimating the optimal market worth of different positions

  — Make the job search more effective & transparent

- Goal: To predict job salary using supervised algorithm

- Dataset: Job description data from UK-based job site

Dataset

- The main dataset would consist of a large number of rows representing individual job ads, and a series of fields about each job ad. The columns are as follows: Id, title, full description, locationRaw, LocationNormalized, contractType, ContractTime, company, Category, SalaryRaw, SalaryNormalized, and SourceName

- Location Tree —This is a supplemental data set that describes the hierarchical relationship between the different normalized locations shown in the job data

ID	TITLE	FULL DESCRIPTION	LOCATION RAW	LOCATION NORMALIZED	CONTRACT TYPE	CONTRACT TIME	COMPANY	CATEGORY	SALARY RAW	SALARY NORMALIZED	SOURCE NAME
12612628	Engineering Systems Analyst	Engineering Systems Analyst … Salary ****K	Dorking, Surrey, Surrey	Dorking		Permanent	Gregory Martin International	Engineering Jobs	20000 - 30000/annum 20-30K	25000	cv-library.co.uk
12612830	Stress Engineer Glasgow	Stress Engineer Glasgow Salary … Glasgow Salary **** to ****	Glasgow, Scotland, Scotland	Glasgow		Permanent	Gregory Martin International	Engineering Jobs	25000 - 35000/annum 25-35K	30000	cv-library.co.uk

Methodology

- Data collection — dataset acquired from a Kaggle competition
- Equipments
  - Development language: Python 3.6.3
  - Integrated Development Environment: Anaconda Navigator
  - Libraries and packages: NumPy, Pandas, Scikit-learn, pickle, features

Methodology (cont.)

- Algorithm: Random Forest — supervised algorithm
- Data manipulation: read dataset -> fit data using final estimator -> save classifier to file
- Prediction: the *predict.py* file is run, which loads the classifier, makes the predictions for each job ID, and writes the *SalaryNormalized* column to a file.

Program

- Program files:
  - data — read dataset & write output to file
  - features — clean the dataset & pick the features to be used
  - train — train the model
  - predict — predict salaries based on test data
- Lab content used in the project — pandas, sklearn, numpy

Algorithm

- Random Forest — a supervised algorithm
- Why?
  - Can be used for both classification & regression tasks
  - Can handle missing values
- sklearn.ensemble.RandomForestRegressor

Results

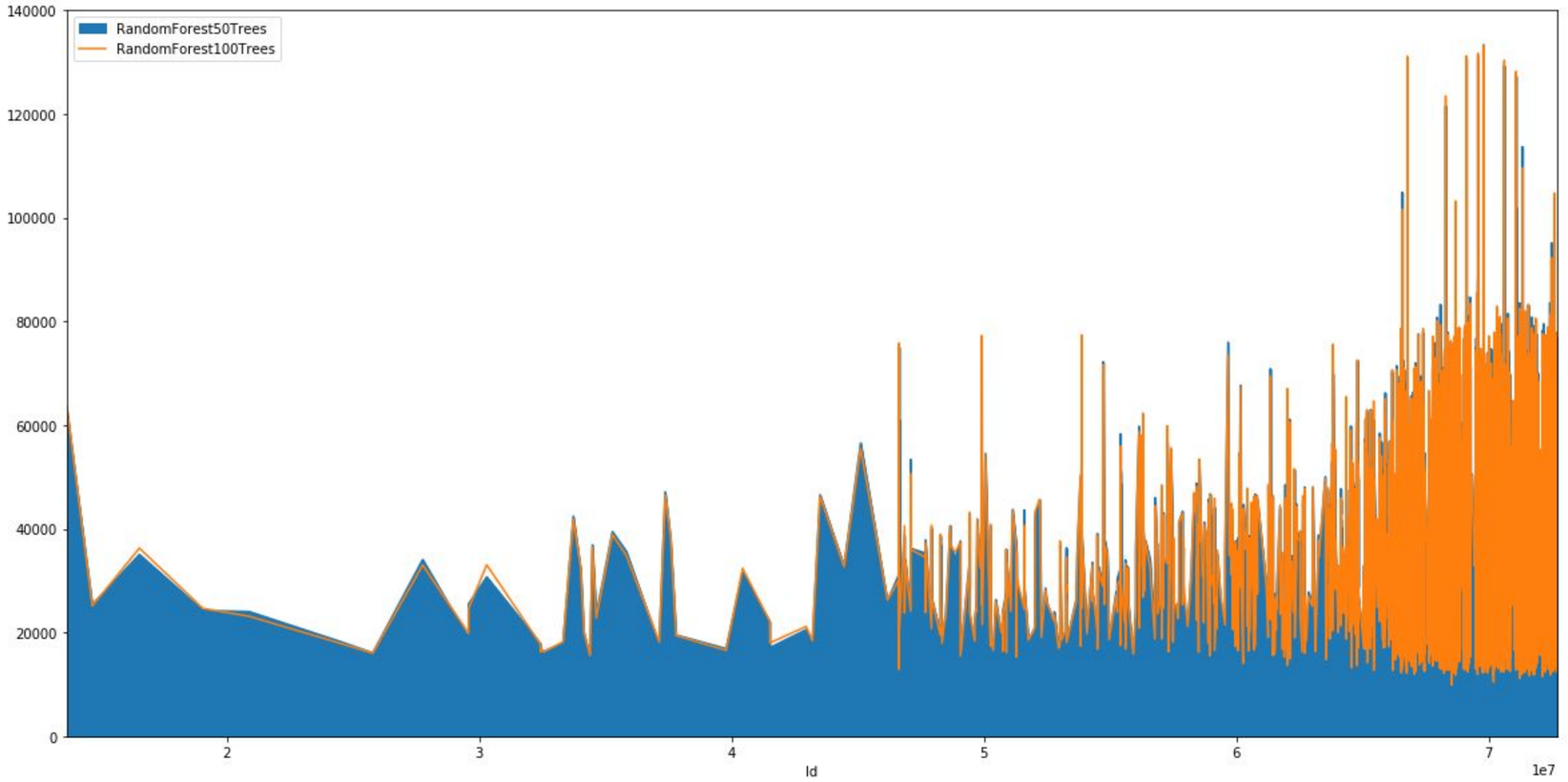

Fig. 1. Result comparison between 50 trees and 100 trees

Results (cont.)

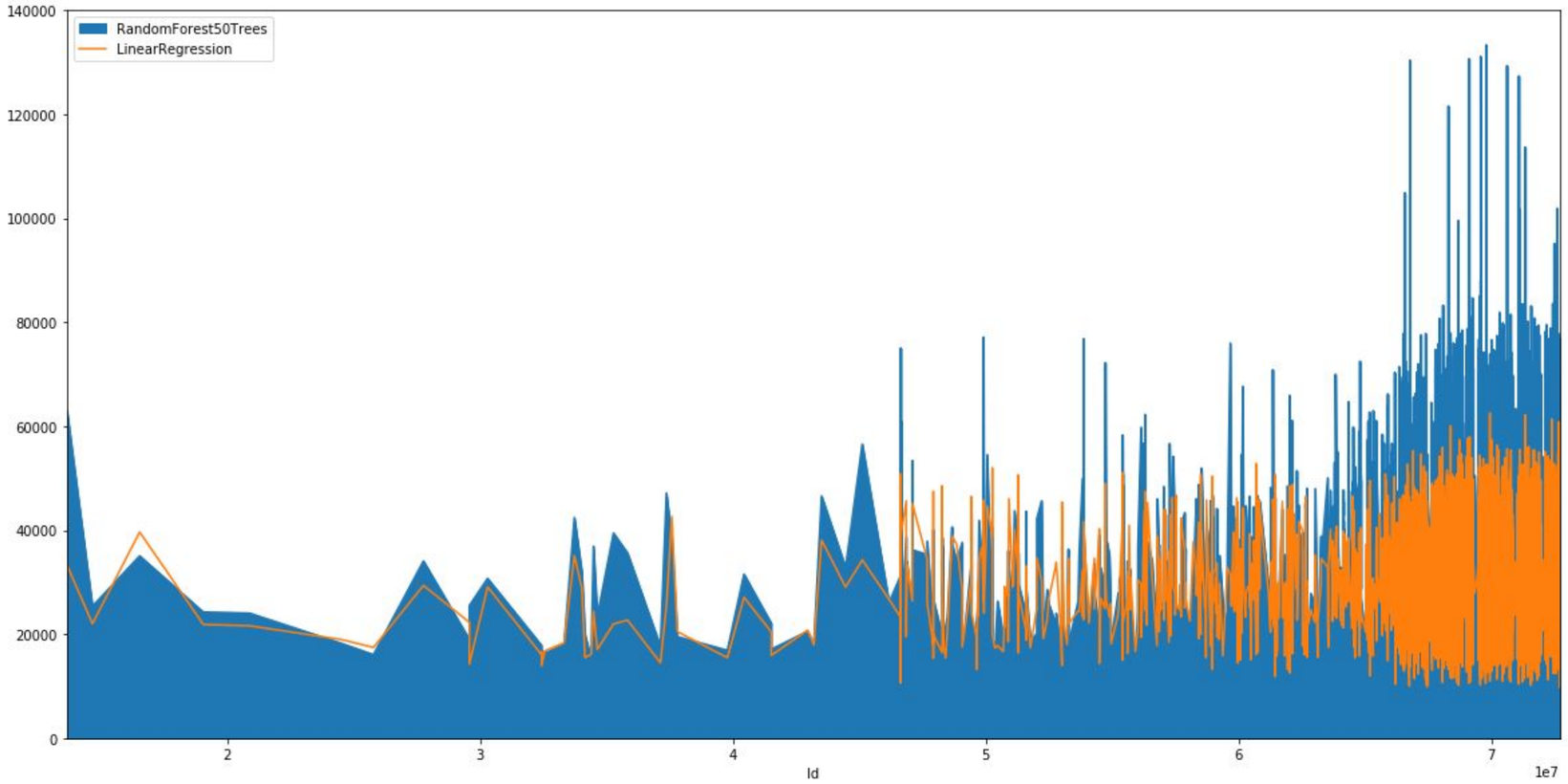

Fig. 1. Result comparison between 50 trees and 100 trees

Conclusion

- Described our iteration of Job Salary Prediction using the dataset from a Kaggle-hosted competition
- This prediction engine can successfully generate a normalized predicted salary based on different criteria from job ads
- Demonstrated the effect on results using twice the number of trees to generate the forest
- Further plans include turning this project into a full-fledged web app by integrating the engine with Django

Acknowledgement

We would like to thank Dr. Iftekharul Mobin, Assistant Professor and Dr. Mohammad Zavid Parvez, Assistant Professor at the Department of Computer Science and Engineering, BRAC University, for their continuous support, guidance, and thoughtful feedback throughout the development lifecycle of this project.

Thank you