Salary Predictor

- Naveenkumar Sivasubramanian (A20378088)
- Sai Ayshwarya Lakshmi Ravichandran (A20378306)

Abstract and Introduction

- The Goal of the project is to build different models which could predict the salary range for a employee and summarize the best model which fits for this problem. This model can be used as a backend for the chat bot.
- We have decided to use classification and regression algorithm with different parameter tuning and cross validation.

Data used

- This dataset is a listing of all current City of Chicago employees, complete with full names, departments, positions, and annual salaries taken from the city of Chicago portal.
- This dataset is used for the training the model with salary as its dependent variable and other attributes as the independent variable.
- The department and position are considered as the questions for the chat bot and the salary range being its result.
- A part of our dataset is separated into the test set for calculating the accuracy.

Column: 4

Records: 32698

Approaches Used

Pre-processing:

- NAN objects in the dataset are removed and the special character for the salary variable is trimmed.
- Now for the best prediction result the continuous variables are converted into categorical variables by creation of bins based on salary value, with a total of four bins present in our model.
- Label encoding is done to convert the text into its corresponding float value.

Fitting the Model:

- We will fit three models (Decision Tree, Random Forest classifier and Logistic regression) for our dataset and summarize the result thereby getting the best model for our project.
- Rather than evaluating the result based on test and train set we have decided to use cross validation technique with 10 folds for the evaluation of result.
- Parameter tuning is applied.

Approaches Used

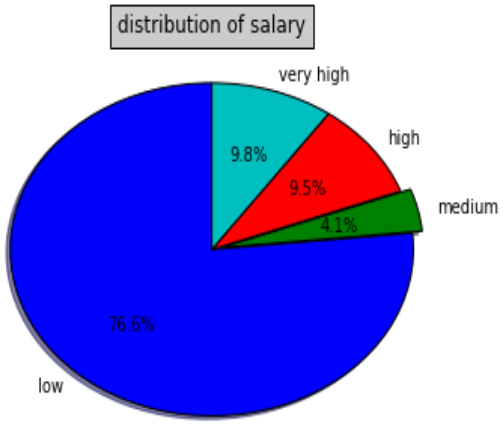
Balancing:

- The major problem with our dataset is the imbalanced property; when the class weight for a particular label is high when compared to others then the result inclines to that class label for most of the prediction.
- The above is solved by three methods and the balanced dataset is fed into the model.
- Classifier parameter class_weight = “balanced” as an argument is an inbuilt balancing method in scikit-learn which adjust the class weight with the class frequency.
- UnderSampling technique.
- SMOTE Analysis (Synthetic Minority Oversampling Technique) which produces more generalized result when considered to random oversampling.

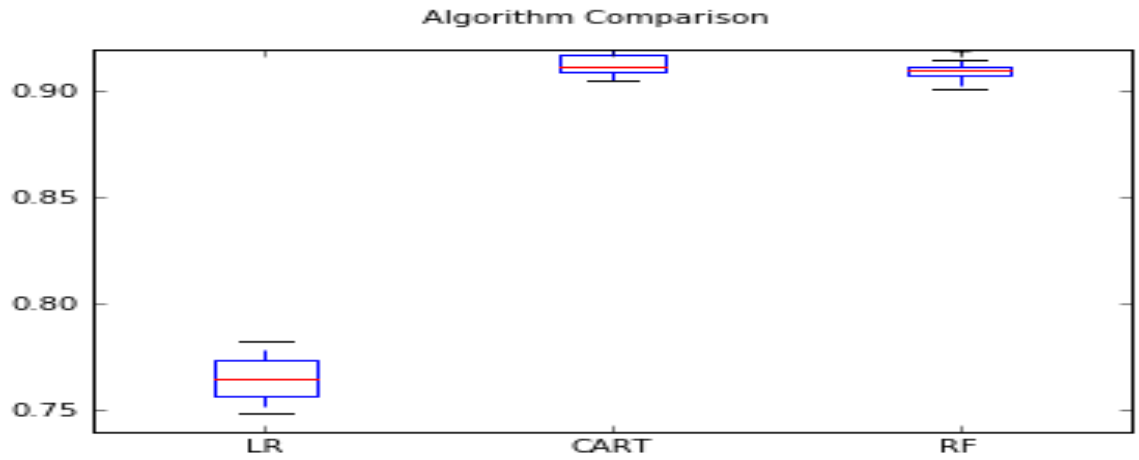
Evaluation:

- Since the data set is imbalance only accuracy cannot be taken as the single best measure.
- So we will evaluate the result with different measure like Mean Absolute error, Mean Square error and average accuracy (1/2*((TP/TP+FN) + (TN/TN+FP))).

Experiments and Result

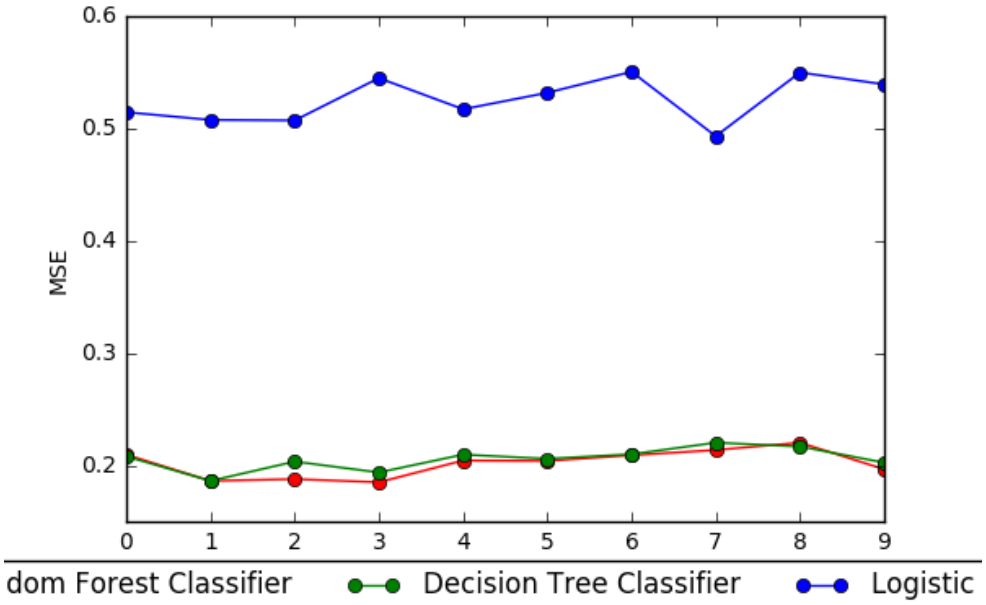


Classifier	Accuracy	Precision	Recall	MSE
Random Forest	0.909	0.912	0.912	0.502
Logistic Regression	0.765	0.586	0.765	1.304
Decision Tree	0.862	0.888	0.862	0.740



Experiments and results

Classifiers	min_samples_split (20)	max_iter (1000)	Criterion (entropy)	Accuracy	Precision	Recall	MSE
Random Forest	Y	N	Y	0.910	0.913	0.912	0.478
Logistic Regression	N	Y	N	0.777	0.789	0.778	1.231
Decision Tree	Y	N	Y	0.891	0.887	0.851	0.509



Conclusion

- Binning and Grouping for analysing continuous class variables.
- Analysis of different classification algorithms for real-time dataset and performance measures.
- Different approaches such Under sampling, SMOTE analysis were used to build a model for imbalanced dataset.
- Evaluation technique for imbalanced datasets.
- This best result can be extended as the model for the chat bot which can include XML or AML as its front end.

References:

- [1] Job Salary Prediction [https://cseweb.ucsd.edu/~jmcauley/cse190/reports/sp15/012.pdf]
- [2] Learning from Imbalanced Data
- [https://www.cs.utah.edu/~piyush/teaching/ImbalancedLearning.pdf]
- [3] SMOTE: Synthetic Minority Over-sampling Technique[https://www.jair.org/media/953/live-953-2037-jair.pdf]
- [4] Model-based clustering and model selection for binned data. [https://tel.archives-ouvertes.fr/tel-01142358/document]
- [5] A STATISTICAL SIGNIFICANCE TEST FOR PERSON AUTHENTICATION [http://www.isca-speech.org/archive_open/archive_papers/odyssey_04/ody4_237]