Wage Prediction of US Permanent Visa applicants

CMPE 255 Data Mining
Fall 2018
Professor: Dr. David

BY

Lu Tang (SJSU ID 011489127)
Niraj Nilesh Dharamshi (SJSU ID 012551266)
Tarang Dhulkotia (SJSU ID 012531129)

1.0 Introduction

We aim to predict wages of US permanent visa applicants using dataset released by US department of labour.

Our main motivation for project was to design a system which can predict future wages given the employer and job type. Since foreign workers have been on the rise for the past couple of decades in US, we started out to explore various datasets where we can find trend of wages for alien workers as this class of labour greatly impacts wage dynamics in the market in US.And thereby,we narrowed our search to H1B datasets and Permanent Visa dataset. Since H1B Visas are allotted for short period and predicting wages for applicants who are going to be in US temporarily didn't meet our target of designing system, we opted for Permanent Visa dataset.

Intended system could be used by policy makers to understand wages of the labour in the near future market and create program to train local college students in high demand-high wage fields. Also our system might be of use to employer who could learn from models prediction about potential cost in the human resource hiring if they depend on foreign labour largely. Lastly future applicants would want to know if they want to stay and work in US, how much wages they should be earning. Our report is organized according to template requirements provided for the project report.

We have formulated wage prediction problem as classification problem in the first approach and regression problem in the second approach. We applied regression models like SGD, Bayesian Ridge etc to understand predictive behaviour of respective models. Evaluation of the same models gave considerable results.Improved predictions were seen when classifier algorithms like KNN were applied. More details on both the approaches are found in the following section.

2.0 System Design & Implementation details

2.1 Algorithms(s) considered/selected

We considered below approaches for our project

- Regression Model based wage prediction.

As dataset is large & feature distribution is not known we thought Decision Tree & Bayesian Ridge models would be appropriate.For simplicity we used Linear Regression Model.

- Classification Model based wage prediction

Decision Tree also works for classification tasks and hence we used it for classification approach too.We thought ensemble method like Random Forest might improve accuracy for the given dataset. Hence we used it.We thought to use KNN too for its simplicity in implementation and understanding.

2.2 Technologies & Tools used

Technologies/Languages:

- Python for Machine Learning Model & Data Mining tasks

Python is used in this project due to its wide community support and library function infrastructure. Also Python offers data mining oriented data structures for easy use.

Tools:

- Google Colaboratory as Jupyter notebook environment

Google Colab is an online shareable notebook where we could work parallely.Also it uses server resources making it best option for collaborative work in resourceless way.

We have used below Python libraries in this project:

- SKLearn

SKlearn library is used for Data Pre Processing, Feature Selection, Machine Learning Algorithms. As it is easy to use and has robust community support.

- Seaborn

Seaborn library is used for Visualising data through different angles.It has powerful visualization function to show complicated data relations.

- Matlib plot

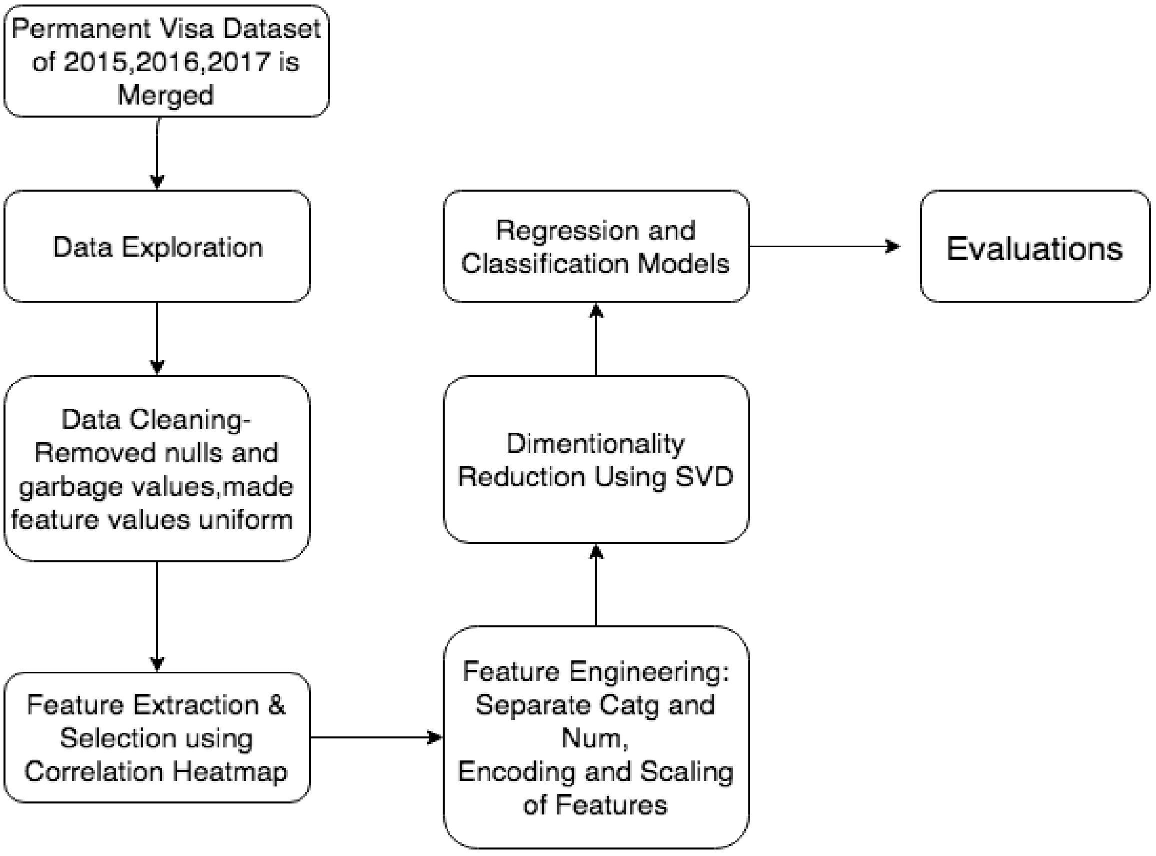
For plotting basic charts and graphs we have used this library.

- SciPy

For representing sparse data and processing sparse data, we have used this library.

2.3 System Design:

Below diagram illustrates our approach 1 and 2:


Permanent Visa Dataset of 2015,2016,2017 is Merged
Data Exploration
Data Cleaning- Removed nulls and garbage values,made feature values uniform
Feature Extraction & Selection using Correlation Heatmap
Feature Engineering: Separate Catg and Num, Encoding and Scaling of Features
Dimentionality Reduction Using SVD
Regression and Classification Models
Evaluations

System Design

Our system consists of the following modules which generally is the flow of Data Mining Task.

1. Data Collection & Merging

Datasets are in three separate csv. They are read into dataframe and merged with some processing steps.

2. Data Exploration

Data is analysed to understand wage distribution, employer-state relationship etc.

3. Data Cleaning

Data is cleaned by removing rows & columns having large null values, outliers are deleted. Columns with more scope of text processing are removed, special characters in data are removed too.

4. Feature Extraction & Selection

Columns are selected based on the correlation heatmap.The columns which had very high correlation were removed, specifically more than 0.6. Further column selection was based on our understanding of the dataset.

5. Feature Engineering

Some of the columns are extracted from existing columns, e.g. column 'employer_name' is converted into column 'application_count_per_employer' by calculating application per employers which gives more meaningful feature than only name. Encoded categorical columns. Standardized numeric columns separately and merged them.

6. Dimensionality Reduction

Due to one hot encoding we got more than 10000 columns, so we applied Singular Value Decomposition on CSR matrix and came up with most important 150 components which reflects 92% variance.

7. Application of Models

Applied many relevant Regression and Classification models on the dataset to predict the actual numeric value of wage as well as the class which denotes wage range.

8. Evaluation

Compared regression models using Means Absolute Error and Classification models using F1-score/Accuracy

3.0 Experiments/Proof of concept evaluation

3.1 Dataset

As mentioned in our introduction, we used dataset named "Permanent Visa". The dataset exists on Enigma Public Portal and can be viewed by clicking below link:

https://public.enigma.com/browse/tag/immigration

Dataset is created and released by United States Department of Labor.

We have used Permanent Visa Applications data for three years. 2015, 2016, and 2017

- Permanent Visa Applications-2015 consists of 88,994 samples and 126 features

- Permanent Visa Applications-2016 consists of 126,143 samples and 126 features
- Permanent Visa Applications-2017 consists of 97,603 samples and 125 features

We have merged all the datasets and worked on consolidated dataset. Dataset has exhaustive list of features related to visa application, employer, applicant, education, profession, prevailing wages and much more. Dataset contains more categorical value columns than numeric value columns.

3.2 Data Preprocessing

At first, we renamed some of the columns in all three the dataframes to make column names uniform before merging. Moreover, in dataset of 2015 we converted state column values from full name to state codes.

We found approximately 50 features which had more than 20 % of missing values. This amounts to around 60000 samples roughly, which can't be imputed so we removed these columns . We used a Python library called missingno to compute the missing values. We decided to drop these columns. After removing columns there were approximately 30000 rows which had null values in one of the columns which we removed.

Columns wage_offer_from_9089 and pw_amount_9089 had wages in many different units so we converted all the wages in to $/year and changed wages accordingly. Moreover, some of rows in wage_offer_from_9089 didn't had not numeric values which we removed.

We also found some interesting observations from data exploration that can be summarized as follows:

- Approx 150,000 cases are Certified-Expired which is almost as status Certified.
- Country with highest number of visa status as Certified - India.
- Top Job having highest approval - Software Development.
- Employer with highest amount of applicants - Cognizant Technology Solutions

3.3 Methodology

Approach-1

We considered prediction of wage as Regression problem here. The dataset majorly contains categorical features so we had to encode all of them using One Hot encoding. One Hot encoding generally gives large feature set if you have unique values in columns. This could make data very high dimensional. So we tried to limit input feature set to be the most relevant and important set. We came up with this feature set from feature correlation heatmap as well as logical & intuitive way as to which features are directly related to applicant's wage.

From final feature set we seperated categorical features and numerical features. We encode Categorical features with One Hot encoding and we standardized numerical features using Min Max Scaler. We merged both CSR matrix. Due to high dimensionality of CSR matrix, we performed SVD on the CSR matrix and captured the 92% variance with 150 components.

We split the data into 90% training and 10% testing set to run the machine learning models. We also did 10-Fold Cross validation while running our Regression models.

We ran below Regression models for this approach to predict wage

- Linear Model
- Lasso Regression
- Ridge Regression
- Decision Tree
- K Nearest Neighbors
- SGD Classifiers

Approach-2

In the first Approach we tried to predict the exact amount of wage which is in the range of 15000 through 500,000. The models we used for Regression didn’t perform extremely well over all the algorithms. So to get the idea of applicant’s wage we thought of making the classes of salary range and create the Classification Problem in lieu of Regression Problem.

We created salary bins according to two criteria

- Equal Frequency - where we kept equal samples in all the salary range. Approximately 58,000 applicants fall under each class.
- Equal width - where we kept salary range as equal for all the class which is $100,000

We tried below Classification models for this approach and predicted wage class

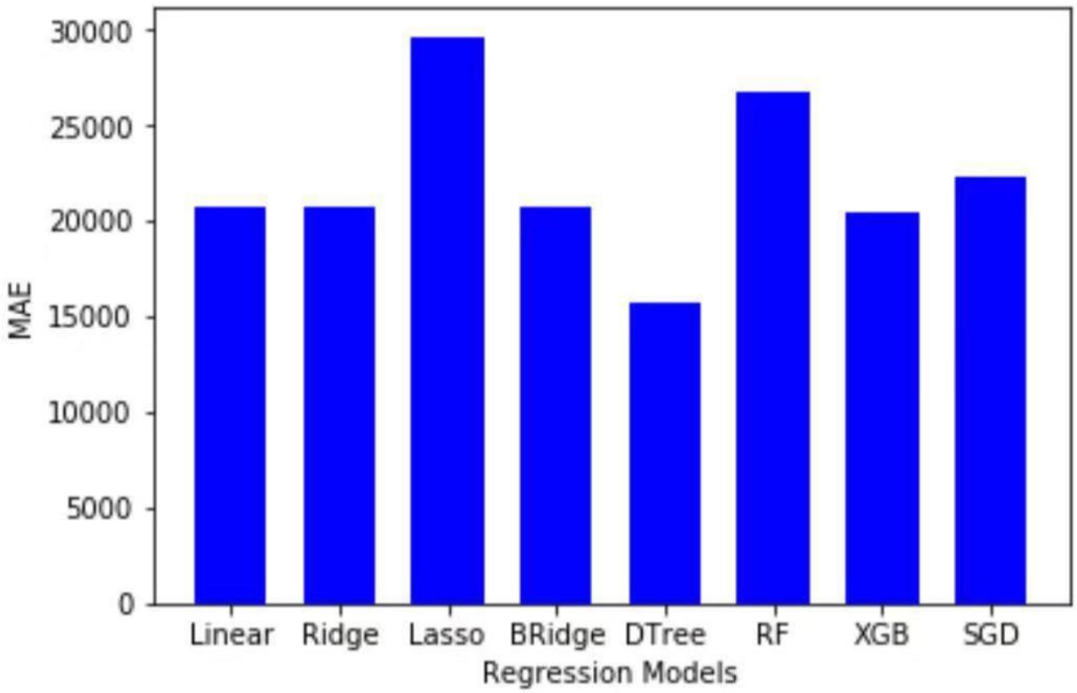
- Random Forest
- Linear SVC
- Gaussian Naive Bayes
- Decision Tree
- K Nearest Neighbors
- SGD Classifiers

3.4 Analysis of results

Evaluation of Approach - 1

We used Mean Absolute Error as metric for evaluating the regression model. Below is the comparison of all the regression models that we performed on our datasets. As evident from the graph - Decision Tree Regressor performed better compared to all other algorithms which gave MAE of around $15,000. We calculated RMSE value for all the models and respective values can be found in the notebook. Best RMSE among all models was observed for Decision Tree Regressor which is approximately $27,000. We also calculated R2 scores for each model. We got best value for Decision Tree which is 0.6.

We tried 10-Fold cross validation on all of these same algorithms but 10-Fold cross gave on an average 1.5 times more error.



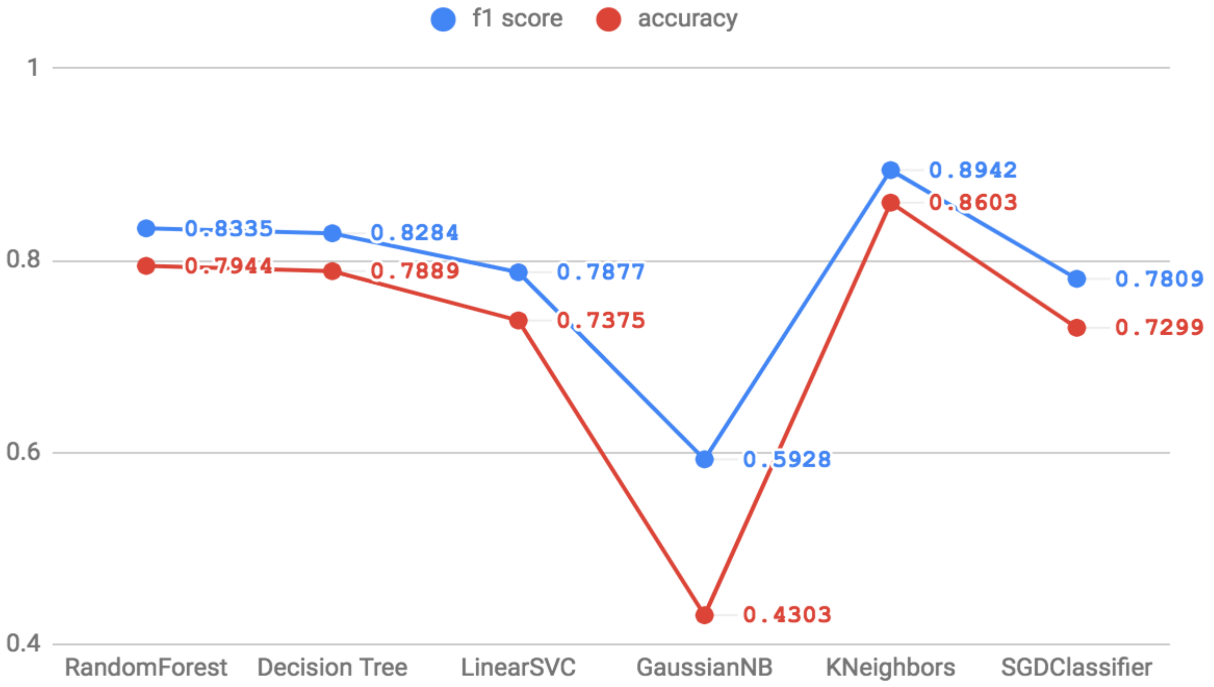
Evaluation of Approach - 2

This is a classification problem so we used Accuracy and F1-Score as evaluation metric for this approach. The graph below shows a good comparison among all the classifiers which we implemented. K Nearest Neighbor Classifier proves to be the best compared to others.

F1 Score of KNN = 0.8942

Accuracy of KNN = 86.03%

Algorithms	KNN	Decision Tree	Random Forest	Linear SVC	SGD	NB
F1-Score	0.8942	0.8284	0.8335	0.7787	0.7809	0.5928
Accuracy	0.8603	0.7889	0.7944	0.7375	0.7299	0.4303



4.0 Discussion & Conclusions

4.1 Decisions made

We made below decision in the process of making data and feature more representative of predicting wage.

- Only those applicants are considered who are certified, as wage details of rejected are not predictive of future behaviour of wages.
- Created new feature called 'application_count_per_employer' to avoid spelling mistakes & its effects in further tasks.
- Removed applicants whose employee number is more than two as applicant himself is the employee and employer which is considered as outliers.
- Removing outliers-whose wage is less than 15000 and more than 500000 as it may affect prediction.

4.2 Difficulties faced

The dataset contained lots of null values and inconsistent datatype. There were many features which contained outliers which we had to remove before processing. Many features seemed redundant for predicting wage. Number of categorical variables in the dataset were quite high so running a regression model seemed like a challenge because after encoding all the categorical features, number of features skyrocketed which we had to reduce using SVD.

During evaluation, 10-Fold cross validation took very long for some of the Regression and Classification models.

4.3 Things that worked well

Data Exploration on unclean data is navigated smoothly. We found some clever implementation on data analysis in Kaggle and taking inspiration from them we obtained good results. Data pre processing and cleaning required maximum effort but paid off at the end for the better classification accuracy.

Classification approach worked better compared to Regression approach in terms of appropriate evaluation metrics.Feature Selection proved most important and helpful component in the project as we tried training models with more features but evaluation metrics deteriorated.

4.4 Things that didn’t work well

- In the first approach, clustering algorithms did not help identify classes as per wage data. Hence we artificially created classes on equal width approach(binning) on wage.
- In the second approach, many features which had negative correlation added noise (reduced MAE) and we could not find any reason for such behavior.
- 10-fold cross validation greatly increased errors across all the models in Regression approach and we could not understand the exact reason.

4.5 Conclusion

Now, It is indeed possible to predict wages based on fraction of attributes of applicants.We have successfully implemented prediction system on two different approaches. Evaluation of respective approaches have reflected sound results on their corresponding metrics.In terms of utility, lawmakers & visa sponsors may find this model useful. Particularly for Policy makers this model may be of great use as it helps them analyse current wage scenario with possible future scenario.Knowledge backed on high confidence models may help them take appropriate decision for future generations. One of the achievement of the project is navigating through highly corrupt and noiseful data.Future improvements to our prediction system that we have considered are:

- Creating more features which could capture better future behaviour of wages.
- More feature preprocessing may help include more columns into model training.
- Prediction using Neural Model may help achieve better accuracy.
- As we had limited time, we could not fine tune Regression models and hence we would like to tune those models again for better performance.

5.0 Project Plan / Task Distribution

Below is the task break up structure in our team.

Research datasets and approaches	All
Data exploration	Niraj & Lu
Data Cleaning	Lu & Tarang
Feature Selection & Engineering	Niraj & Tarang
Algorithm research	All
Report Writing	All
Evaluation	All
Classification Approach	All(2 Algos each)
Regression Approach	All(2 Algos each)
Approach Implementation	All
Presentation Slides	Lu

References:

1. ‘US Permanent Visa Applications_v1.1’ Kaggle.[Online]. Available:https://www.kaggle.com/elzawie/us-permanent-visa-applications-v1-1.Accessed:11-22-2018
2. ‘Predicting the Outcome of H-1B Visa Application’ Stanford CS229.[Online]. Available:http://cs229.stanford.edu/proj2017/final-reports/5208701.pdf.Accessed:11-20-2018.

Note:

As our system can work with the full dataset, we could not create subset of our dataset for your evaluation. However, we can provide the entire dataset through Google Drive.