Assignment 1  
Regression Models

short line

Ronik Jayakumar

Student ID: 24680264   
29/03/2024

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Executive Summary 2**](#_xvtyf97iwmd0)

[**2. Business Understanding 3**](#_if9iwijz2pg3)

[a. Business Use Cases 3](#_bnjfuh5ewogw)

[**3. Data Understanding 4**](#_yg03y3w5hig9)

[**4. Data Preparation 5**](#_tj77rrmixzhi)

[**5. Modeling 6**](#_6vwh14nc44ga)

[a. Approach 1 6](#_uv8sn0ik6017)

[b. Approach 2 6](#_wtwxuxemqtvc)

[c. Approach 3 6](#_nrl8g3ri789x)

[**6. Evaluation 8**](#_mngiyx5tet6w)

[a. Evaluation Metrics 8](#_qx20m6hrk8uo)

[b. Results and Analysis 8](#_25v57zsj5m7g)

[c. Business Impact and Benefits 8](#_hkob76wu4d6q)

[d. Data Privacy and Ethical Concerns 9](#_uf1z6gbsejg6)

[**7. Deployment 10**](#_8rqux13i2o01)

[**8. Conclusion 11**](#_9hr11g79asdx)

[**9. References 12**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

The specific business use case in our project is for a Talent Placement Cell specializing in generating job placements for young graduates or individuals. The features areL

* Offering pay estimates based on candidates' credentials and industry norms
* Improving placement Success by providing accurate wage forecasts
* Maintain good relationships with employers as well as candidates

Challenges and Opportunities:

* Keeping up with market trends and industry standards is critical for making accurate compensation estimates that reflect current job market trends
* Helping candidates stand out in a competitive job market
* Ensuring candidate happiness with appropriate job placements and corresponding salaries.
* Strict authorization standards need to be setup by an organization when dealing with salary information.

Machine Learning Key Role:

Machine learning algorithms are important in this scenario as they:

* Analyze Diverse Data Sources to generate salary estimates for candidates.
* Improve placement success by giving data driven insights
* Help mitigate biases that exist in humane decisions.
* Offer scalability and flexibility allowing the talent placement cell to handle large volumes of candidates and job data.

1. Key Objectives

**Key objectives of the project:**

* Salary Prediction: The goal of the project is to create a machine learning model that can predict salary based on various background information available about the candidate.
* Data driven decision making: The project aims to enable data driven decision making in talent placement leading to more informed candidate selection and job placements.

Stakeholders and their Requirements:

* Candidates expect accurate salary predictions and personalized job recommendations based on their skills.
* Employers require access to qualified job candidates that match their required description to choose from.
* Educational institutions aim to enhance their alumni placement rates by realistic salary expectations and training them in relevant job skills.

Addressing Stakeholder Requirements:

* Address the need of a reliable salary estimator by analyzing historical data
* Sophisticated job matching algorithms use applicant traits to find viable matches that fit the needs of both candidates and employers.
* Automating applicant screening, resume parsing, and interview scheduling duties thereby increasing overall efficiency.

Use of CRISP-DM methodology:

All the above will be implemented by using the CRISP-DM methodology. The 6 phases are:

* Business Understanding
* Data Understanding
* Data preparation
* Modelling
* Evaluation
* Deployment [4]
* short dash

# Data Understanding

The Dataset used for the project consisted of the following features:

* There was a total of 3 datasets: - Training, Validation, Testing
* All the three datasets consisted of 31 columns. The training dataset consisted of 2998 rows whereas the validation and testing sets consisted of 599 rows each.

Data collection methods and Limitations:

* The data was provided to us by the University and the collection methods remain unknown. Limitations may include possible biases such as underrepresentation.

Variables and features available in the Dataset:

* The dataset consists of 27 Integer and Float columns, 3 columns with objects and 1 column with dates of birth. One of the integer columns named ‘ID’ is the unique identifier column for the 3 datasets.
* Relevant features of a graduate such as marks, degree, specialization, aptitude, personality, and a slight glimpse into the person with age and gender is received.

Exploratory Data Analysis:

* EDA methods such as summary statistics, data visualization (e.g., histograms, heatmaps), and correlation analysis were used to analyze variable distribution and relationships.
* EDA assisted in identifying outliers, missing values, and trends in the data, which then informed preprocessing procedures and model selection decisions.

short dash

# Data Preparation

Observations:

* Data exploration to understand datatypes, presence of null values, mean and median values.
* Target variable exploration to understand what our current salary column looks like. In this dataset, the salary column portrayed many outliers due to a large spectrum of specializations and skills covered by the dataset.
* Duplicate values from the validation and test datasets existed in the training dataset. This would cause the model to overfit by having to learn and predict the same data.

Preparation:

* The date of birth column was converted into an age column of type integer .
* The top 10 features minus 12 graduation was identified using a correlation matrix
* The p values of the 3 object columns corresponding to the target column was calculated which showed specialization has the highest correlation. This column was encoded using Label Encoder.
* Box plots of all numerical features were generated to check for outliers., we have decided to clip the data to the upper and lower bounds due to the limited availability of data.
* The computer programming column had a total of 650 values corresponding to -1 which indicates a lack of skill which is replaced by its mean.
* The overall dataset including the target variable is scaled using Standard Scaler.
* The reasoning for also scaling the target value is to avoid erratic RMSE values due to the sheer difference in numbers. Salary ranges through a large set of numbers which maybe a difficult RMSE score to comprehend.
* Finally, the entire training set is split into training, validation, and testing. The split is generated at 80% training, 10% validation, and 10% testing except for KNN where the split was 70 – 15 -- 15.

short dash

# Modeling

A total of 5 different models were used in this project. Given below is an explanation of each.

Experiment 1 (Multivariate Linear Regression):

* Multiple Linear regression is a statistical technique used to analyze the relationship between two or more independent variables with a dependent variable.[1]

Experiment 2 (Ridge Regularization):

* The ridge regression model can manage multicollinearity while also preventing overfitting by adding a regularization factor to the linear regression equation. [2]
* Parameter tuning is done by using the K Fold method to find the optimum value of alpha.

Experiment 3 (Lasso Regularization):

* Lasso regression is a regularization technique used over regression methods for a more accurate prediction using shrinkage.[2]
* Parameter tuning is done by using the K Fold method to find the optimum value of alpha.

Experiment 4 (ElasticNet Regularization):

* Elastic net is a combination of the two most popular regularized variants of linear regression: ridge and lasso. [3]
* Parameter tuning is done to both L1 and L2 parameters, that is alpha and L1 ratio.

Experiment 5 (K Nearest Neighbors):

* The K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values.
* Parameter tuning is done to find the optimum value for K to get the best RMSE scores.

short dash

# Evaluation

## Results and Analysis

Experiment 1 (Linear Regression):

* Baseline model:
  + Training RMSE – 1.0
  + Validation RMSE – 1.0
  + Testing RMSE – 0.96
* Multivariate Linear Regression:
  + Training RMSE – 0.872
  + Validation RMSE – 0.868
  + Testing RMSE – 0.843
* The model learnt the training set quite well and continued to give good predictions over the validation and test sets. This model is capable of giving predictions of salaries.

Experiment 2 (Ridge Regularization):

* Baseline model:
  + Training RMSE – 1.0
  + Validation RMSE – 1.0
  + Testing RMSE – 0.96
* For the ridge regression model, we first calculate the value of alpha using the K Folds method, In this case, it is calculated that the best value of alpha is 20.
* Ridge model:
  + Training RMSE – 0.872
  + Validation RMSE – 0.868
  + Testing RMSE – 0.843
* The model learnt the training set quite well and continued to give good predictions over the validation and test sets. This model can give good predictions of salaries.

Experiment 3 (Lasso Regularization):

* Baseline model:
  + Training RMSE – 1.004
  + Validation RMSE – 1.001
* We then calculate the value of alpha using the K Folds method, In this case, it is calculated that the best value of alpha is 0.001
* Lasso model:
  + Training RMSE – 0.872
  + Validation RMSE – 0.868
  + Testing RMSE – 0.843
* The model learnt the training set quite well and continued to give good predictions over the validation and test sets. This model can give good predictions of salaries.

Experiment 4 (ElasticNet Regularization):

* Baseline model:
  + Training RMSE – 1.0
  + Validation RMSE – 1.0
  + Testing RMSE – 0.96
* For the ElasticNet regularization model, we first calculate the value of alpha and L1. In this case, it is calculated that the best value of alpha = 0.01 and L1 = 0.1.
* ElasticNet
  + Training RMSE – 0.872
  + Validation RMSE – 0.868
  + Testing RMSE – 0.843
* The model learnt the training set quite well and continued to give good predictions over the validation and test sets. This model can give good predictions of salaries.

Experiment 5 (K Nearest Neighbors):

* The baseline model consists of a KNN model with K = 5. The results were as follows:
  + Training RMSE – 0.78
  + Validation RMSE – 0.992
* We then calculate the best value of K using the K Folds method, In this case, it is calculated that the best value of K is 49
* KNN:
  + Training RMSE – 0.867
  + Validation RMSE – 0.921
  + Testing RMSE – 0.816
* We see a slight overfit in the validation set but the overall model performs well in the unsee test set.

Overall, we see the linear regression model doing the best as compared to KNN.

## Business Impact and Benefits

* Improved wage estimation accuracy can help with talent placement and negotiation.
* A better knowledge of the elements that influence compensation may help organizations optimize workforce management strategies.
* Accurate compensation predictions promotes openness and justice in employment procedures, promoting confidence between employers and employees.
* The overall candidate satisfaction levels remain high.

## Data Privacy and Ethical Concerns

* Data privacy implications: Maintaining the confidentiality and security of sensitive personal information used in wage prediction models
* Ethical considerations: Fairness, openness, and accountability in model creation and deployment,
* Steps Taken: Implementing privacy-preserving strategies such as data anonymization, access limits, and encryption can help protect sensitive information. Furthermore, implementing ethical reviews and audits throughout the project's lifespan aids in the identification and resolution of potential biases.

short dash

# Conclusion

* Linear regression along with its regularizations proved to have the best RMSE scores with all the regularization techniques giving very similar scores.
* The performance of KNN varied with choice of hyperparameters K which highlights the importance of optimization.

.

Reflection:

* The research effectively met its aim of creating realistic wage prediction models.
* The stakeholder's expectations for transparent, fair, and accurate wage calculation were addressed.
* The project addressed data privacy and ethical concerns by using privacy-preserving mechanisms.

Future Work and Recommendations:

* Experimenting with sophisticated machine learning approaches like gradient boosting and neural networks
* Exploring other variables or data sources, such as employment positions, industrial sectors, or geographical considerations
* Collaboration with Indigenous communities and stakeholders to include Indigenous perspectives and values
* Models must be updated and refined on a regular basis based on stakeholder feedback and evolving job market trends

short dash

# References

[1] : Multiple Linear Regression: Definition, Example, and Applications – Mayur Badole – 17th Jan 2024 – Analytics Vidhya - <https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/>

[2]: [1]: Ridge and Lasso Regression in Python | Complete Tutorial (Updated 2024) - Aarshay Jain - 11 Jan, 2024 - Analytics Vidhya - <https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-completetutorial/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters>

[3]: ElasticNet Regression: The Best of L1 and L2 Norm Penalties – Aditya Kakde – Nov 11 2023 – Medium - <https://adityakakde.medium.com/elastic-net-regression-the-best-of-l1-and-l2-norm-penalties-7a340e2387d6#:~:text=While%20ridge%20employs%20an%20L2,both%20L2%20and%20L1%20penalties>.

[4]: What is CRISP DM? – Nick Hotz – March 26th 2024 – Data Science Process Alliance - https://datascience-pm.com/crisp-dm-2/

short dash