MIE1624 Assignment 2 Daire Crawford
March 12th, 2023 1002225745

Prediction Salary Bracket Based on 2022 Kaggle Survey Data

Question 1: Data Cleaning and Feature Engineering

The focus of the data cleaning was converting categorical variables to numerical variables, removing redundant or irrelevant variables, and removing data that has significant missing values. Questions relating to the respondents personal preferences (e.g., media consumption or platform ease of use) were removed as they were considered out of scope of the model objective and not expected to have a large influence in the prediction. A detailed review of the data cleaning process can be seen in Appendix A.

Many questions were already categorized into binary responses. For example Question 13_1 through 13_14 contained the name of the IDE specific to that variable (e.g., Q13_1 corresponded to the use of JupyterLab) if the respondent utilized that IDE in their work, NaN otherwise. It is assumed that if the respondent utilized that IDE that the data was captured by the non-NaN values in that variable and therefore non-NaN values were converted to '1' to represent use of that specific IDE, and NaN values were converted to 0 if the IDE is not used. This method was performed for a total of 16 variables (Appendix A).

Categorical features were either one-hot encoded if they contained no inherent order to their values (e.g., Q23, Job Title) or converted to ordinal numeric variables if there was an inherent order to the data (e.g., Q2, Age).

In addition to encoding the data, certain features were engineered to provide a coarser level of information. For example, Q4, relating to the individual's country of residence, was converted to three separate categories – "USA", "Developed_NonUSA", and "Developing". This was done in lieu of removing countries with few respondents in an effort to retain as much information from the data as possible despite some questions having few respondents (e.g., the majority of individuals who filled out the survey were from the USA and India). Additionally, it is expected that the respondent's country of residence will have a significant influence on the model and by retaining the data in some capacity leave avenues for further research into the more granular influence of country possible for future studies. Similar feature engineering techniques were applied to a total of seven survey questions, details of which can be seen in Appendix A.

Question 2: Exploratory Data Analysis and Feature Selection

The cleaned dataset contained 8017 rows and 216 variables. Due to the large size of the dataset, feature selection was performed prior to intensive data exploration to provide more relevant information relating to the features in the model. Feature selection was performed using the *SelectKBest* function from Sklearn. *SelectKBest* allows one to select the top *k* features that have the largest contribution to the target variable, which is determined via ANOVA F_1 scores or Chi-Squared test for classification tasks. Utilizing the F_1 score metric, the top 10 features from the dataset were collected, which can be seen in Appendix B.

Figure 1 shows the correlation matrix of the top 10 variables. The matrix indicates that the "region_usa", and "exp_encoded" variables are positively correlated with the target variable, while "region_developing" is negatively correlated. Trends between experience, age, and salary are also plotted in Figures 3 & 4 in

Daire Crawford
1002225745

Appendix C. Additional variables also demonstrated some correlation, such as the use of ethical machine learning techniques and GPU and GBM use, while individuals with more years of experience were more likely to utilize these tools. Values for the mode, median, and standard deviations of the features can be seen in Appendix D.

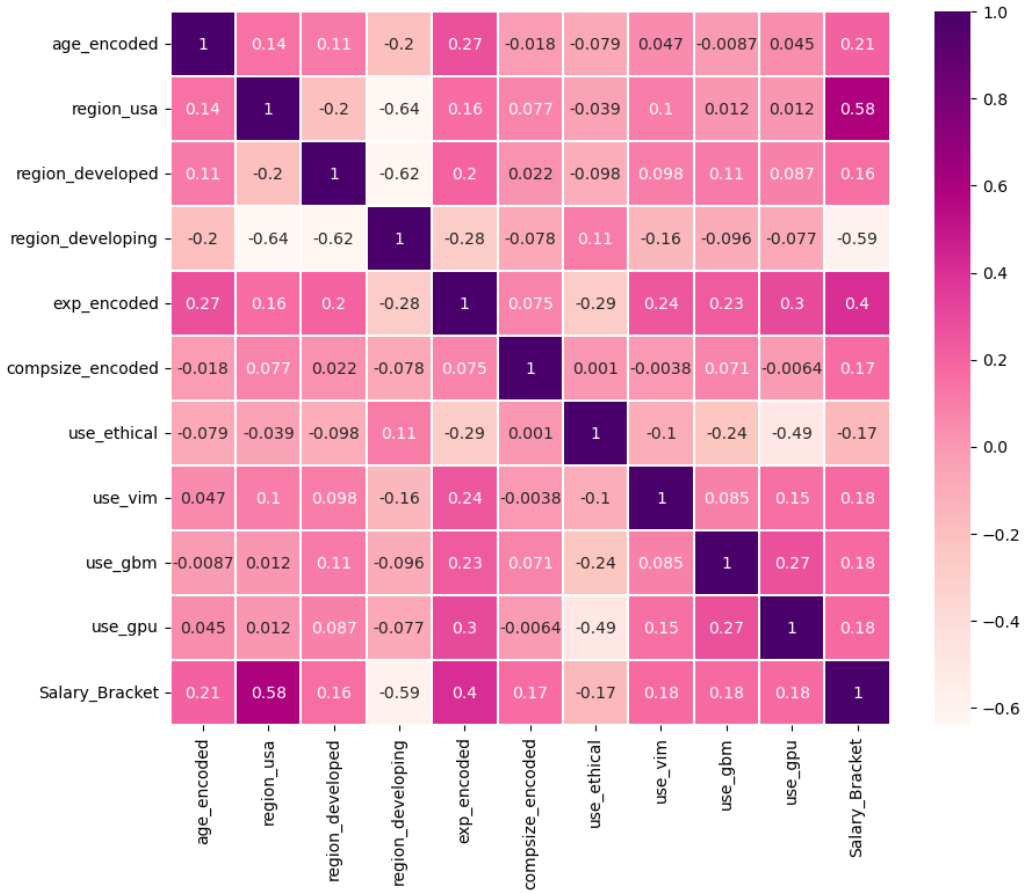

Figure 1: Correlation matrix of the top 10 variables in the dataset.

Question 3 & 4: Model Implementation + Tuning

The initial model implementation was run without any hyperparameter tuning and had an accuracy of 0.376 and a small variance of approximately $2x10^{-4}$. When hyperparameter tuning was implemented the accuracy was marginally better at 0.378. The model with the best accuracy had a 'C' value of 0.5 and an 'L2' penalty applied. This application of the L2 penalty vs no penalty likely increased the bias of the model as the L2 penalty applies regularization to the feature weights, causing some features to be heavily considered in the model and others less so, meaning that there would be high bias and low variance. This could explain the low variance value achieved in all models. In addition to this, scaling of the model features was not performed. This was not performed as the majority of the features are one-hot encoded (binary) , but upon review, likely should have in order to better normalize between the binary and ordinal features. This likely also impacted the model bias/variance as the features as the features that contain a higher range are more likely to influence the calculations between the distance of data points, essentially biasing the model to the features with larger magnitudes and reducing variance [1].

Daire Crawford
1002225745

Overall model accuracy is not a suitable method for this problem as there is a large variation in the potential outputs and the ordinal nature of the outputs (i.e., 14 ranked classes). If the model erroneously predicts a class of 13 when the true class is 14 the accuracy will be lowered the same amount as when the model predicts class 1 when the true class is 14. While the model did incorrectly classify the input in both cases, the scale of the inaccuracy is not equal (class 13 is significantly closer to class 14 than class 1), therefore you would want a model performance metric that would reflect this.

Question 5: Testing & Discussion

The model performs similarly in the training and with-held test set, with an accuracy of 0.378 and 0.376 respectively. This model is likely heavily biased and the random sampling of the data was performed in an effort to mitigate this. However, this can be further improved by scaling the data points, or by reducing the number of classes, as each class would have a larger number of data points and therefore increase the variance in the data of each class. Additionally, selecting different hyperparameters to tune, or different values to utilize in the grid search, may also reduce the bias. Currently the model is underfitting as it has high bias and low variance. This underfitting because the model is utilizing few features that are heavily weighted to predict the class outcome (Figure 2).

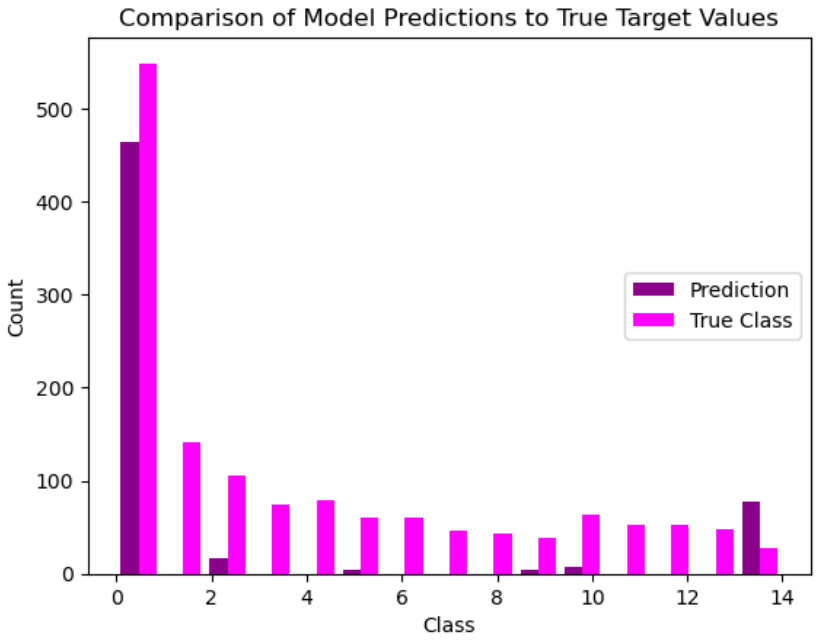

Figure 2: Trend of the respondents age group in relation to average salary for that experience bracket.

References

[1] B. Roy, "All about Feature Scaling," *Medium*, Feb. 04, 2023. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35 (accessed Mar. 12, 2023).

[2] "THE EFFECT OF PROFITABILITY, FIRM SIZE, LEVERAGE AND FIRM VALUE ON INCOME SMOOTHING by ajhssr.editor - Issuu," Jan. 31, 2023. https://issuu.com/ajhssr.editor/docs/k227018994 (accessed Mar. 12, 2023).

Daire Crawford
1002225745

Appendix

Appendix A

Table 1: Overview of data cleaning process for each variable.

Question Number	Alterations	Justification
Q2: Age	Converted to ordinal bins	Required for input into regression, age has a natural order that allows it to be converted to ordinal
Q3: Gender	Removed all responses except Man/Woman, converted to binary	Responses outside of Man/Woman responses were minimal and wouldn't contribute to the model very much
Q4: Country	All responses were kept and altered to USA, Non-USA Developed, and Developing Removed, "other" and "do not wish to disclose" responses	The majority of responses are from USA and India, but it's expected that Country would have an influence on overall salary, so I aimed to keep the data and engineer it into a larger scope "Other" and "I do not wish to disclose" responses were minimal and not able to be categorized into the new features so those responses were removed.
Q5: Currently a Student?	Removed, all responses are the same, "No".	Since all the responses are the same there is no variation in which the model can use for prediction and its an unnecessary additional variable
Q6: Learning Platforms	Converted to "university" and "Online" features, binary	If this variable was one-hot encoded the vast majority of the responses would be '0' and the level of information would be too granular to gain useful information. By engineering the features into two main categories it may provide insights into the use of online or traditional schooling and its

		impact on salary, and provides the opportunity for future research if this is an important feature in the model.
Q7: Helpful platforms	Removed	The perceived usefulness of different learning platforms is subjective and would not provide useful information for the model
Q8: Education level	Converted to ordinal	Prefer not to say lumped with no highschool edu
Q9: Research Publications	Removed	Originally converted to binary then found that 2898 entries were “NA” -- more than 10% rule of thumb
Q10: ML Research	Converted to binary	
Q11: Programming Experience	Converted to ordinal variable	Years of experience has a natural order that allows it to be converted to ordinal
Q12: Programming Languages	Converted to # of languages known	If this variable was one-hot encoded the vast majority of the responses would be ‘0’ and the level of information would be too granular to gain useful information. By engineering the features into the number of languages known it may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Q13: IDEs	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q14: HostedNotebooks	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q15: DataVisLibraries	Converted to binary	Single column only contains one categorical value, NaNs were

		converted to 0.
Q16: MLExperience	Convert to ordinal variable	Years of experience has a natural order that allows it to be converted to ordinal.
Q17: MLFrameworks	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q18:MLAlgorithms	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q19:ComputerVision	Converted to binary variable 0 = Dont use CV 1 = Use of CV	If this variable was one-hot encoded the vast majority of the responses would be '0' and the level of information would be too granular to gain useful information. By engineering the features into whether or not the individual uses computer vision it may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Q20:NLP	Converted to binary variable 0 = Dont use NLP 1 = Use of NLP	If this variable was one-hot encoded the vast majority of the responses would be '0' and the level of information would be too granular to gain useful information. By engineering the features into whether or not the individual uses NLP may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Converted to binary	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q22:MLHubs	One-Hot Encoded	Single column contained multiple categorical values.

Q23:Title	One-Hot Encoded	Single column contained multiple categorical values.
Q24: Industry	One-Hot Encoded	Single column contained multiple categorical values.
Q25: CompanySize	Converted to Ordinal	Company size may affect salary, as larger companies may afford larger salaries [2].
Q26:DSEmployees	Removed	Data formatting in excel file was incorrect, saved as a date format.
Q27:MLUse	Removed	This information is captured through other questions relating to ML use, title, and industry, ultimately making it repetitive
Q28:Activities	Removed	This information is captured through other questions relating to ML use, title, and industry, ultimately making it repetitive
Q29:Salary	Removed	Duplicate Column after original salary values were converted to ordinal
Q30:MLSpending	Removed	Not within scope, personal spending may correlate with salary but in reality likely will have little influence on the model
Q31:CCPlatforms	Converted to binary 0 = no use of cloud computing 1 =use of cloud computing	If this variable was one-hot encoded the vast majority of the responses would be ‘0’ and the level of information would be too granular to gain useful information. By engineering the features into whether or not the individual uses cloud computing may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Q32:PlatformPreference	Removed	Preference of Platform is out of scope and wouldn’t relate to

		salary
Q33:CCProducts	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q34:DataStorage	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q35: DataProducts	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q36: BITools	Converted to binary 0 = no use of BI Tools 1=Use of BI tools	If this variable was one-hot encoded the vast majority of the responses would be '0' and the level of information would be too granular to gain useful information. By engineering the features into whether or not the individual uses BI Tools may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Q37:ManagedML	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q38:AutoML	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q39:MLServe	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q40: MLMonitor	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q41:Ethical AI	Converted to Binary	If this variable was one-hot encoded the vast majority of the responses would be '0' and the level of information would be too granular to gain useful

		information. By engineering the features into whether or not the individual uses ethical AI tools may provide insights into the breadth of knowledge and relation to salary, and provides the opportunity for future research if this is an important feature in the model.
Q42:Hardware	Converted to binary	Single column only contains one categorical value, NaNs were converted to 0.
Q43:TPU	Converted to ordinal	Number of times an individual has used TPUs has a natural order allowing it to be converted to an ordinal variable
Q44:Media	Removed	Preferences for media sources will likely not have a impact on overall salary

Appendix B

Table 2: Overview of the top ten features used in the dataset

Feature Name	Description
age_encoded	Ordinal variable with each age represented in the data
region_usa	Binary variable representing if the respondent is working in the USA
region_developed	Binary variable representing if the respondent is working in a developed nation outside of USA
region_developing	Binary variable representing if the respondent is working in a developing nation
exp_encoded	Ordinal variable correlating to how many years of programming experience the respondent has
compsize_encoded	Ordinal variable representing the size of the company the respondent works at
use_ethical	Binary variable representing if the respondent utilizes ethical AI techniques in their work
use_vim	Binary variable representing if the respondent

	utilizes VIM in their work
use_gbm	Binary variable representing if the respondent utilizes gradient boosted models in their work
use_gpu	Binary variable representing if the respondent utilizes ethical GPUs in their work

Appendix C

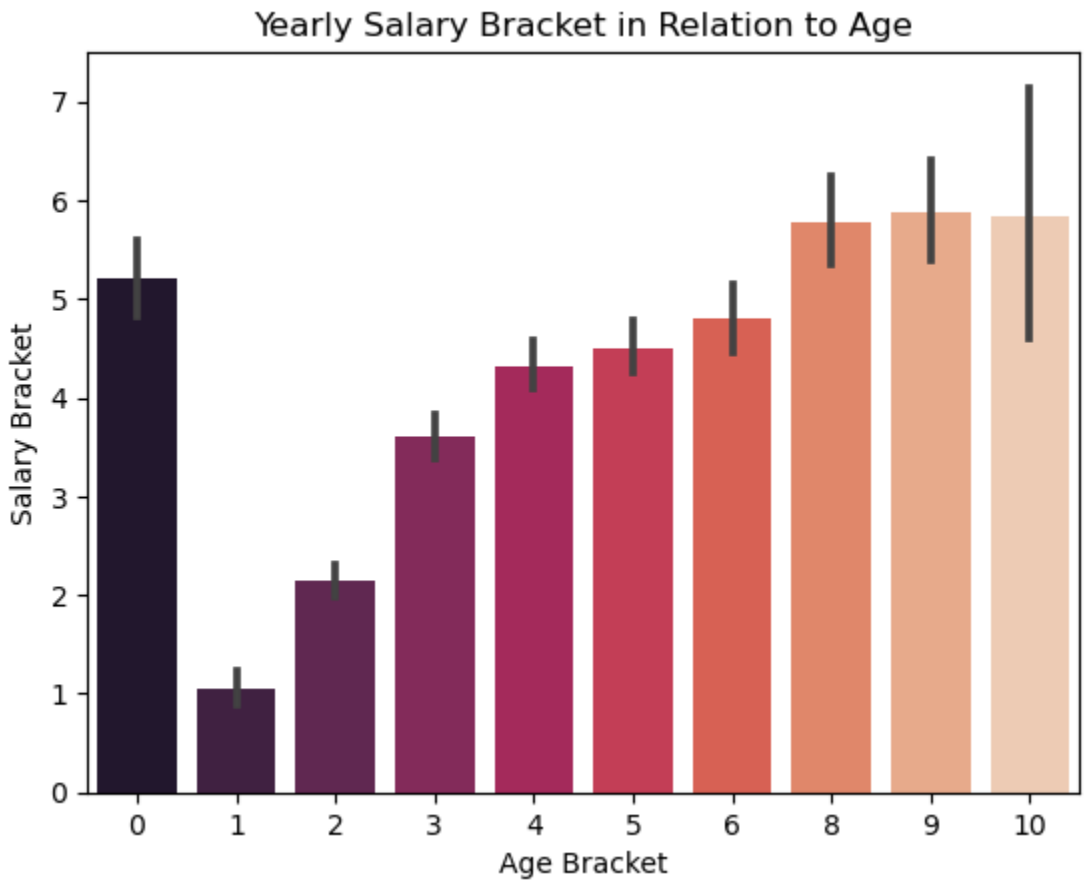

Figure 3: Trend of the respondents age group in relation to average salary for that age bracket.

Daire Crawford
1002225745

Figure 4: Trend of the respondents age group in relation to average salary for that experience bracket.

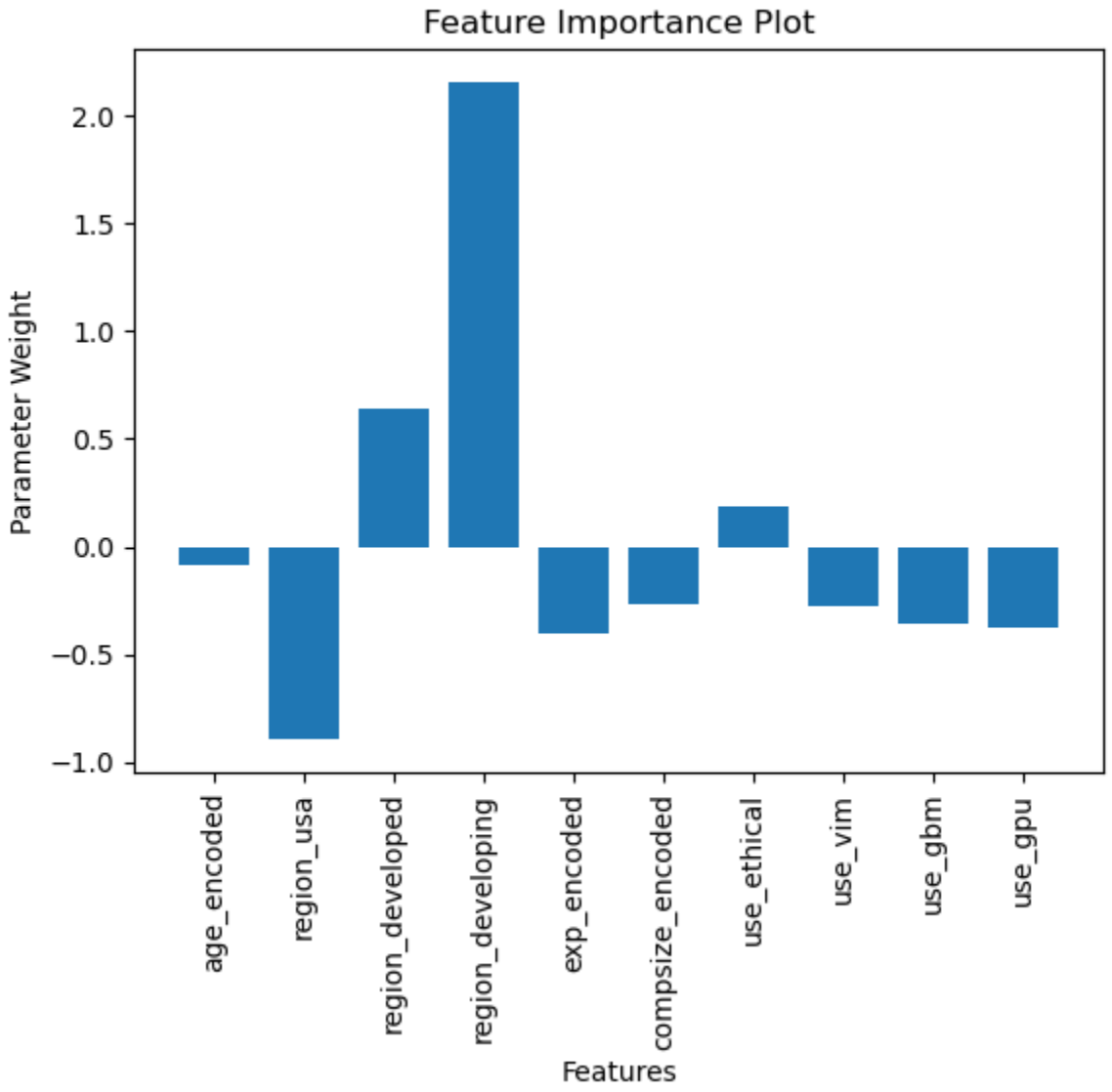

Figure 5: Feature parameter weights for the top 10 features used in the model.

Appendix D

Table 3: Mode, Median, and Standard Deviation of the top 10 features

Feature	Mode	Median	Standard Deviation
age_encoded	2	3	2.3
region_usa	0	0	0.37
region_developed	0	0	0.37

region_developing	1	1	0.47
exp_encoded	2	3	1.84
compsize_encoded	4	2	1.50
use_ethical	1	1	0.50
use_vim	0	0	0.29
use_gbm	0	0	0.47
use_gpu	0	0	0.47