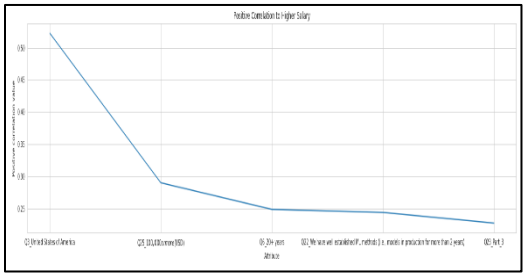
Question 1: Data Cleaning

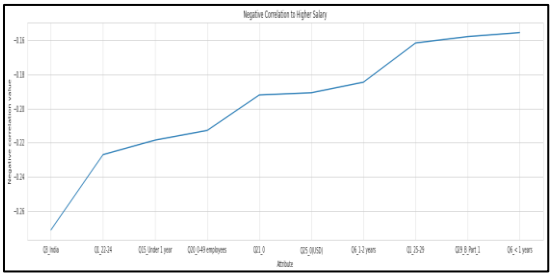
The data contains many null values, by performing a count of null values in the data, I can see which columns to clean and handle them. I focused on columns without multiple parts since the data will be separated into separate columns using dummy encoding. 'Q8', 'Q11', 'Q13', 'Q15', 'Q25', 'Q30', 'Q32', 'Q38' have null values. Q8, 11, and 13 had their missing values filled using the mode because one of the values dominated the column. Q15 has the value of ‘Under 1 year’ and ‘Not using ML’ may be similar based on how you interpret the question. Since Q25 doesn’t have many null values (~1-2% of the 10,000 data points), I dropped the null values. I dropped columns 30, 32, and 38 because they contain too many missing values.

I then converted the categorical data into numerical data using dummy encoding. Dummy encoding allows us to create a new column for each unique value of a feature, depending on the column, this will create many extra columns. The dummy variable sets the value as 1 and all other values as 0 for that column. This encoding was also used because there are questions that are answered as multiple parts and placed into different columns (Ex. Q7 Part 1, etc.). By using this method, I do not have to find a way to add up the columns or fill their missing values. By converting the data into numerical data, it is easier to work and process with.

Question 2: Exploratory Data Analysis and Feature Selection

There are multiple methods on feature selection and determining the most important features. After cleaning the data, I used the correlation values between the salary buckets and the encoded features to see which features are the most important on determining a higher salary. From the positive correlation values, we can see that residing in the US has the highest correlation to a higher salary. The two next leading features are how much their company spent on ML, and their experience in coding. From the negative correlation values, we can observe the features that do not provide the highest salary. We see that residing in India has the highest negative correlation to a higher salary, the next two leading factors are the participant’s age and experience (or lack of experience) in coding. The country where the survey participant resides and their experience in coding are at a first observation, the most important factors for a higher salary. From the top 100 features that are positively correlated to a higher salary and top 100 features that are negatively correlated to a higher salary, I’ve combined the features manually to train the model and to reduce the number of features we need to work with.





These 2 figures show the top 5 features that are positively and negative correlated to a higher salary respectively. Feature engineering is a useful tool in machine learning because it is important which features are used to train the model, if we use the wrong features, the model will rely on the wrong features to make its predictions. Dummy encoding is a feature engineering technique used to separate the unique values into different columns which makes it easier to process. How we handle missing values is also a feature engineering technique that may affect our model.

Question 3: Model Implementation

I split the dataset into training and test data, in which we do not touch the test data and will use it at the end to validate our model. The salary range is divided into 15 buckets so 14 binary classification is required to separate them. A 10-fold cross validation is applied and the model accuracy and F1 score is calculated for each bucket. The model's F1 score varies a bit between folds and doesn't seem to follow a specific trend. The average model accuracy gets better increases as the bucket range increases and the variance decreases. The average model salary increases from 77.78% to 98.48% between buckets 0 to 14, and the standard deviation decreases from 1.84% down to 0.537% between buckets 0 to 14. However, The F1 score tends to decrease as we move up in salary buckets, in this dataset, F1 score is a more useful number representation because it seems that the data is skewed to low salaries. For example, if the model says all feature will provide a low salary (dominant class), it is more likely than not will provide true positive and true negative results used in the accuracy calculation. This may be also the bias-variance trade-off, as we move to higher buckets, it seems to be overfitting the training sample to the lower salary target values. We will see if it's overfitting if it provides a good score for the training data but not the test data. Scaling/ normalization of the features aren't necessary because they are encoded as 1's and 0's using dummy encoding.

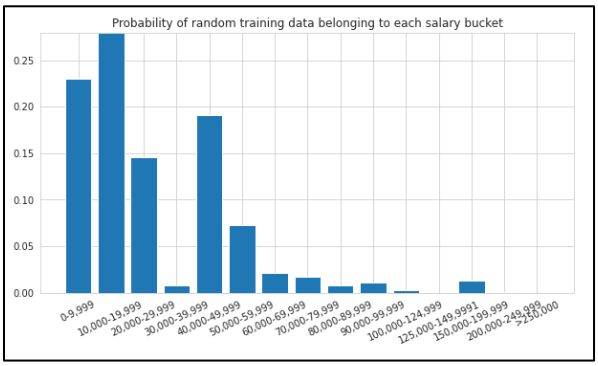
Question 4: Model Tuning

A hyperparameter used in my ordinal logistic regression mode was the norm used in the penalization. There are multiple solvers however, some are not used such as Stochastic Average Gradient (sag), which is impractical for large N (because it remembers the most recently computed values for approximately all gradients) and Liblinear is used in 1 vs rest scenarios. I used newton-cg and lbfgs which uses a L2 norm penalty. Another hyperparameter used was C, the inverse of regularizations strength, which is the coefficient of the cost function for the regularized logistic regression and a range from 0.1 to 10 was set. The number of max iterations is also a hyperparameter, however, to avoid divergence and long computations time, a limit of 500 iterations was set.

The metric used to select the optimal hyperparameter and model was the F1 score because for our case, our ordinal binary classification is fit on skewed data. The number of counts between 0 and 1 is very different, especially at higher salary brackets since the lower salary bracket is the dominant class. I mentioned the difference between accuracy and F1 score in question 3. Using F1 score not only tells us which predictions we have right (accuracy), it penalizes us against the ones we also got wrong (precision and recall). If out F1 score is close to 1, this mean that our prediction matches our target value, if our F1 score is close to 0, this means that our model isn't working very well.

For all 14 binary classifications, the best model coefficient and intercept is stored and used to solve for the probability (which inverse of regularization value and which solver provides the best F1 score). The probability of each survey participant belonging to each salary bracket is calculated using the logistic function. Since the brackets are cumulative, the bracket's probability is subtracted from each other to get that specific salary range. The sum of the probability for each bracket should be equal to 1, the maximum probability for a bracket is then used to classify which bracket the survey participant belongs to. In the below example, the random participant will fall into salary range of $10,000 –$ 19,999. When

we compare the F1 score between the original default model and the hyperparameter tuned one, the hyperparameter tuning one provides us with a slightly better F1 score.



Question 5: Testing & Discussion

When we apply our optimal model to the test data, we obtain an accuracy of 79% and an F1 score of 0.16. Comparing that to our training data accuracy of 81.4% and F1 score of 0.243 we are slightly overfitting since it works better with training data. One method to increase the accuracy of the test and training set is to use different feature selection methods. By using other feature selection techniques such as PCA rather than selecting the features manually, we may be able to choose features that may be more useful on determining higher salary. Other ways to improve the accuracy is to use different feature engineering methods since dummy encoding sets all the values to 0 and 1 and since the data is skewed, the F1_score will not be as good. Another method is to assign weights to certain features since some features has many more possible choices, the choices will be spread thin among many different columns. There are many multiple ways to improve accuracy of our model, other methods are to use different regression algorithms, choose different hyperparameters to tune, etc. The distribution of the true target variable values and their predictions on both the training set and test set are plotted below.

Our model predicts that the salary range if $0-$9999 is the most frequent salary range. Our model for both training and test data underestimates the number of participants in the other salaries ranges. By trying to balance the data set with those who have a higher salary, more features can be used and extracted to determine the key features. There also seems to be a bump at the $100,00 – 124,999 range which may be the most common salary range for those in machine learning and data analytics, $0-9,999 may be the dominant range because people are uncomfortable sharing their salary in a survey.