# Question Generated by Participants

### Participants generated specific questions about the linear regression model ‘Section 1: Question about the model situation: Is the relationship between the independent variable(s) and the dependent variable strong? OR Is the model credible?’

How reliable is the model predicting house prices based on features like size?

Does the regression model for predicting student performance show a strong correlation?

Is there a strong linear relationship between advertising spending and product sales?

Can we trust the model’s prediction results for housing prices in the Boston dataset?

Is the model credible?

Is the relation betwen input and output variables solid enuf to rely on this model?

How credibly does the model link the number of study hours to exam scores?

Model strong or not? Can it tell us anything real about GDP and education levels?

I want to know if the model built on the auto-mpg dataset is believable.

Does the regression analysis of sugar intake and blood glucose show a meaningful connection?

### Participants generated specific questions about the linear regression model ‘Section 2: Question about how (different) independent variable affects the dependent variable: How does the dependent variable change with an increase in the independent variable(s)?’

As the number of rooms increases, how does the house price change in the Boston Housing dataset?

What happens to the sales revenue if we increase advertising budget by $1,000?

How do different levels of education affect individual income levels?

Wht impact does higher age have on insurance costs?

How does the dependent variable respond when the independent variable rises?

When GDP increases, how does the unemployment rate change in the model?

Does increasing exercise duration lead to a drop in body fat percentage?

If we raise the interest rate, what effect does it have on housing demand?

Hw does inflation impact consumer spending accrdng to the model?

How do changes in temperature affect electricity usage?

### Participants generated specific questions about the linear regression model ‘Section 3: Which independent variable has the greatest impact on the dependent variable?’

Among all features in the housing dataset, which one most strongly influences house prices?

What factor contributes the most to predicting a car’s fuel efficiency?

Which independent variable most heavily impacts student test scores in our model?

Which feature—income, education, or age—affects health outcomes the most?

Wich variable has the highest effect on sales?

Can you tell me which input has the biggst impact on y?

What’s the strongest predictor in our regression of diabetes progression?

What is the top influencing factor on employee attrition rate in the HR dataset?

Out of all independent variables, which one is most critical for predicting electricity consumption?

Which var affects the outcome variable the most in our linear model?

### Participants generated specific questions about the linear regression model ‘Section 4: Question about if the independent variable has a significant effect on the dependent variable: Do the independent variable(s) have a significant effect on the dependent variable?’

Does education level significantly influence annual income in the regression model?

Are any of the predictors in the Boston Housing dataset statistically significant?

Do advertising expenditures significantly affect monthly sales figures?

Are the independent varables meaningful in predicting the target outcome?

Which features are statistically relevant in explaining the variance of fuel efficiency?

Does GDP significantly impact unemployment rate in the model we built?

Do age and gender play a significant role in predicting medical expenses?

Are the effects of predictors on the output variable strong enough to be considered significant?

Can we say that the relationship between income and happiness is statistically valid?

Is there evdence that the inputs have a real effect on prediction results?

### Participants generated specific questions about the linear regression model ‘Section 6: Question about if model have risk of overfitting: Is there a risk of overfitting in the model? OR Does the model have a risk of overfitting?’

Does our regression model for predicting house prices show signs of overfitting?

Is the model trained on the diabetes dataset too closely fitted to the training data?

Are we seeing overfittng in the model’s performance metrics on test data?

Might the model be memorizing noise instead of learning general trends?

I suspect our car price prediction model is overfit—is that true?

Does this model generalize well, or is it possibly over-fitted to the training set?

Could this high R² on training but low on testing mean the model’s overfited?

Is the regression capturing true patterns or just overfitting the training data?

Do the residuals suggest that the model is too complex and might be overfitted?

Is there any evidence that our linear regression model suffers from overfit?

### Participants generated specific questions about the linear regression model ‘Section 1: Question about the model accuracy: How accurate is the model? OR Is the model credible?’

What’s the accuracy of the decision tree classifier on the iris dataset?

Is our spam detection model reliable when applied to new email samples?

How accrate is the model in predicting customer churn?

Does the model classify images correctly most of the time?

Is the performance of the model good enough for deployment in real-time fraud detection?

Can we trust the classification results from this model on unseen data?

Hw well does the model predict loan default?

What’s the model’s precision and recall on the test set?

Is this classifier good enuf to use in real-world applications?

How dependable is the model trained on the breast cancer dataset?

### Participants generated specific questions about the linear regression model ‘Section 2: Question about how different independent variables affect the dependent variable (usually, classification results): Which independent variable(s) positively influence one of the classes of the dependent variable? OR Which independent variable(s) exert a negative impact on one of the classes of the dependent variable?’

Which features increase the likelihood that a customer will cancel their subscription?

In the breast cancer dataset, what factors contribute positively to predicting a malignant tumor?

Which variables are negatively associated with being classified as high risk in loan applications?

Wich predictors boost the chances of a student being placed in the “pass” class?

What independent variables reduce the probability of a patient being classified as diabetic?

Hw do age and income impact the model's classification of customers as VIPs?

Which factors push the model to label a transaction as fraudulent?

Do any variables discourage the model from labeling a product review as positive?

What makes the classifier more likely to categorize an applicant as “accepted”?

Are there any attributes that negatively influence whether a user is predicted to churn?

### Participants generated specific questions about the linear regression model ‘Section 3: Question about the impact of independent variables: What is the most influential factor? OR For each classification result, which independent variable has the greatest impact? OR On average, which independent variable has the greatest impact on classification results?’

Which feature is the most important in classifying whether a tumor is benign or malignant in the breast cancer dataset?

What factor has the strongest effect on whether a customer is predicted to churn?

For our loan approval model, which variable plays the biggest role in determining classification outcomes?

Which input varible contributes most to the model's classification decisions?

On avg, which feature most affects whether an applicant is accepted or rejected?

Can you tell me what variable the classifier relies on most to decide credit risk levels?

What is the top influencing factor across all classes in the image classification task?

Hw do different predictors rank in terms of their impact on classification accuracy?

In the HR dataset, which variable most strongly influences employee attrition classification?

What’s the most significant factor the model uses to predict customer satisfaction levels?

### Participants generated specific questions about the linear regression model ‘Section 4: Question about if model have risk of overfitting: Is there a risk of overfitting in the model? OR Does the model have a risk of overfitting?’

Does our decision tree classifier for email spam detection show signs of overfitting?

Is the classification model for predicting customer churn overfitting to the training data?

Are we seeing any signs that the image classifier is too tightly fitted to the training set?

Do performance metrics suggest the modl is overfit?

Might the model be memorizing training examples rather than generalizing to new ones?

Hw can we know if our model is overfitted on the HR dataset?

Is there a big gap between training and test accuracy indicating overfiting?

Could our model’s high accuracy on training but low on validation mean it’s overfit?

Is the fraud detection model too complex and likely to overfit the dataset?

Do you think the classification model will generalize well, or is it at risk of over-fitting?

### Participants generated specific questions about the linear regression model ‘Section 5: Question about which independent variable has the greatest impact: Which independent variable has the greatest impact on the dependent variable?’

Which predictor most strongly influences whether a customer is labeled as a churn risk?

In the breast cancer dataset, which feature best determines if a tumor is malignant?

What input variable has the greatest effect on predicting if an email is spam or not?

Wich factor is most important in the classification of student performance levels?

Which independent varable contributes the most to determining customer satisfaction categories?

Hw does the model decide who is at high vs low credit risk? What variable matters most?

In our HR attrition model, which feature drives the prediction most significantly?

What single variable most affects whether someone gets approved for a loan?

Is age the most influential factor in predicting diabetes classification outcomes?

Which input does the classifier rely on most to separate the two classes?

### Participants generated specific questions as negtaive samples

What is the format of the dataset, and does it contain any missing values?

How many rows and columns are there in the training set?

Can you show a histogram of the “age” feature in the dataset?

What types of features (categorical vs. numerical) are included in the dataset?

Has the dataset been normalized or scaled before modeling?

What is the correlation matrix among all numeric variables?

How are null values handled in the employee attrition dataset?

Can you list the unique categories in the “education level” column?

Are there any outliers in the “income” feature, and how are they treated?

How is the dataset split between training and testing sets?

What encoding method is used for categorical features like gender and location?

How many classes are there in the target variable?

Can you show the distribution of the dependent variable in the test set?

What are the first five rows of the dataset?

Are any variables strongly correlated with each other in an undesired way?

What is the mean and standard deviation of the “price” feature?

How many missing entries are there in the “bmi” column?

Did you use one-hot encoding or label encoding for the target variable?

What software or programming language was used to process the data?

What’s the overall shape of the dataset after dropping the null values?

Does the model include regularization terms such as L1 or L2?

What kind of loss function was used in training this model?

Can we visualize how each class is separated in the latent space?

Is feature selection performed before model training?

Which features are included in the final trained model?

Does the model architecture include any dropout layers?

Are categorical variables treated differently in this classification pipeline?

Did we consider multicollinearity before fitting the regression model?

Are ensemble methods like bagging or boosting used in this model?

Is the model interpretable by human experts? 1

How are the model’s hyperparameters tuned? 0

What kind of preprocessing is applied before fitting the model? 0

How is feature importance calculated in this model? 3

Is PCA or dimensionality reduction used before classification? 0

How many features are used for prediction after filtering? 0

Are decision boundaries smooth or abrupt in the final model? 0

Was cross-validation used to evaluate the model? 6

Do any variables need transformation before being included in the model? 0

Is this model more complex than a baseline logistic regression? 0

Which training algorithm was applied to build the final model? 0

# Experimental Results

Correctly matching 98 out of 100 positive samples and successfully rejecting 94 out of 100 negative samples.