Review of trivago presentation

1. Duplication of sentences is too much

2. 24 mins for 1st guession is too much

Focus

Q1.

ONLY 2 sentence

Question 2:

1. SCQ framework discussed in 1 min >>>>> 2.45

2. Quick glanced of variable types and indicate hour of day is a circular variable and could be transformed with sine and consine to be included as predictors. >>>>> 4.10

Exploratory analysis – 10 min

Descriptive analysis — 1 min

## Data visualization to ascertain certain assumptions required by some models – 5 min

Viz target and quantitative variables – 1 min

### Visualizing the relationship between target variable and quantitative predictors – 2 min

### Visualization of categorical predictors — 2 min

## Feature selection – 10 min

#### Feature selection for categorical features with low cardinality – 2 min

## Handling predictors with high cardinality – 2 mim

## Categorical Feature engineering – 1 min

## Numeric feature selection – 1 min

#### Visualizing outliers Boxplot to visualize outliers – 1 min

### Handling missing data – 2 min

### Using insights gained from exploratory analysis to inform modelling approach – 1 min

Modelling — 10 min

Baseline model definition – 2 min

Modelling and CV – 1 min

Test evaluation – 1 min

Hyperparameter tuning – 3

Bagging – 2 min

Summary

a)Are all locales impacted by the loss of data equally?

Data quality checks remains an important part of analytics particularly to check data completeness and accuracy. Analyzing the distribution and presence / influence of missing data provides a clearer background and circumstances surrounding the data hence undertaken.

To gain a clear understanding of the question for the analysis, equal data loss is interpreted as the proportion of missing data among locales is homogenous and equal.

Determining this requires a hypothesis test and chi square test of homogeneity is used.

Q2 Review

(i) Do not overstress the SCQA

(ii) Do not talk about log\_scale for visualization