Bryan Camacho
CAP5106

HW1

Q1: In training set, which features are available?

'PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'

Q2: In training set, which features are categorical?

'Name', 'Sex', 'Ticket', 'Cabin', 'Embarked'

Q3: In training set, which features are numerical(e.g., discrete, continuous, or time series based)?

'PassengerId', 'Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare'

Q4: In training set, which features are mixed data types?

Ticket

Q5: In training set, which features contain blank, null or empty values? In test set, which features contain blank, null or empty values?

train columns: ['Age', 'Cabin', 'Embarked']
test columns: ['Age', 'Fare', 'Cabin']

Q6: In training set, what are the data types(e.g., integer, floats or strings)for various features?

PassengerId int64, Survived int64, Pclass int64, Name object, Sex object, Age float64, SibSp int64, Parch int64, Ticket object, Fare float64, Cabin object, Embarked object

Q7: To understand the distribution of numerical feature values across the samples, please list the properties, including count, mean, std, min, 25% percentile, 50% percentile, 75% percentile, max, of numerical features?

SHOWN IN CODE

Q8: To understand the distribution of categorical features, we define: count is the total number of categorical values per column; unique is the total number of unique categorical values per column; top is the most frequent categorical value; freq is the total number of the most frequent categorical value. Please list the properties, including count, unique, top, freq,of categorical features?

SHOWN IN CODE

Q9: Can you observe significant correlation (average survivied ratio>0.5) among the group of Pclass=1and Survived?
Yes, significant correlation is seen.

If Pclass has significant correlation with Survivied, we should include this feature in the predictive model. Based on your computation, will you include this feature in the predictive model?

Yes, I would add this feature to my model.

Q10: Are Women (Sex=female) were more likely to have survived?
Yes

Q11: Let us start by understanding correlations between a numeric feature (Age) and our predictive goal (Survived). A histogram chart is useful for analyzing continuous numerical variables like Age where banding or ranges will help identify useful patterns. The histogram can indicate distribution of samples using automatically defined bins or equally ranged bands. This helps us answer questions relating to specific bands (e.g., infants, old). Please plot the histogramsbetween ages and Survived (Figure1 is an example), and answer the following questions:

•**Do infants (Age <=4) have high survival rate?** High probability for survivability

•**Do oldest passengers (Age = 80) survive?** Most did survive

•**Do large number of 15-25 year olds not survive?** Yes, a large number of 15-25 year olds did not survive.

Based on your analysis of the figures,

•**Should we consider Age in our model training? (If yes, then we should complete the Age feature for null values.)** Yes, we should. Age shows an indicative distribution.

•**Should we should band age groups?** No

Q12: We can combine three features(age, Pclass, and survivied)for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values.

Please plot the plot using python, and answer the following questions:

·Does Pclass=3 havemost passengers, however most did not survive? True

·Do infant passengers in Pclass=2 and Pclass=3 mostly survive? Yes

·Do most passengers in Pclass=1 survive? Yes

·Does Pclass varyin terms ofAge distribution of passengers? Yes, Pclass 1 had older and Pclass 3 had more younger passengers

·Should we consider Pclass for model training? Yes, Pclass is a good predictive feature for survivability.

Q13: We want to correlate categorical features (with non-numeric values) and numeric features. We can consider correlating Embarked (Categorical non-numeric), Sex (Categorical non-numeric), Fare And answer the following questions:

•Do higher fare paying passengers have better survival?
Seem to make a difference in survivability only on Embarked C

•Should we consider banding fare feature?
Yes

Q14: What is the rate of duplicates for the Ticket feature? 681 duplicates which may mean one ticket number was assigned per party.

Is there a correlation between Ticket and survival?
Does not seem to be correlated

Should we drop the Ticket feature?
Yes

Q15: Is the Cabin feature complete? No, it is not

How many null values there are in the Cabin features of the combined dataset of training and test dataset? There are 1014 null values

Should we drop the Cabin feature? Yes, we should drop it. Too much missing data

Q16-20 on code (linked below)

Additional Questions:

·Approximately how many hours did you spend on this assignment? About 3 hours

·**Which aspects of this assignment did you find most challenging? Were there any significant stumbling blocks?** I found the coding of the visualizations the most challenging to fit the structure.

·**Which aspects of this assignment did you like? Is there anything you would have changed?** I enjoyed the questions it asked based on the visualizations. I would have added more open-ended questions for us to look more into the data.

Link to code:
https://github.com/CamachoBry/CAP5610-ML/blob/main/Homeworks/HW1/HW1.ipynb