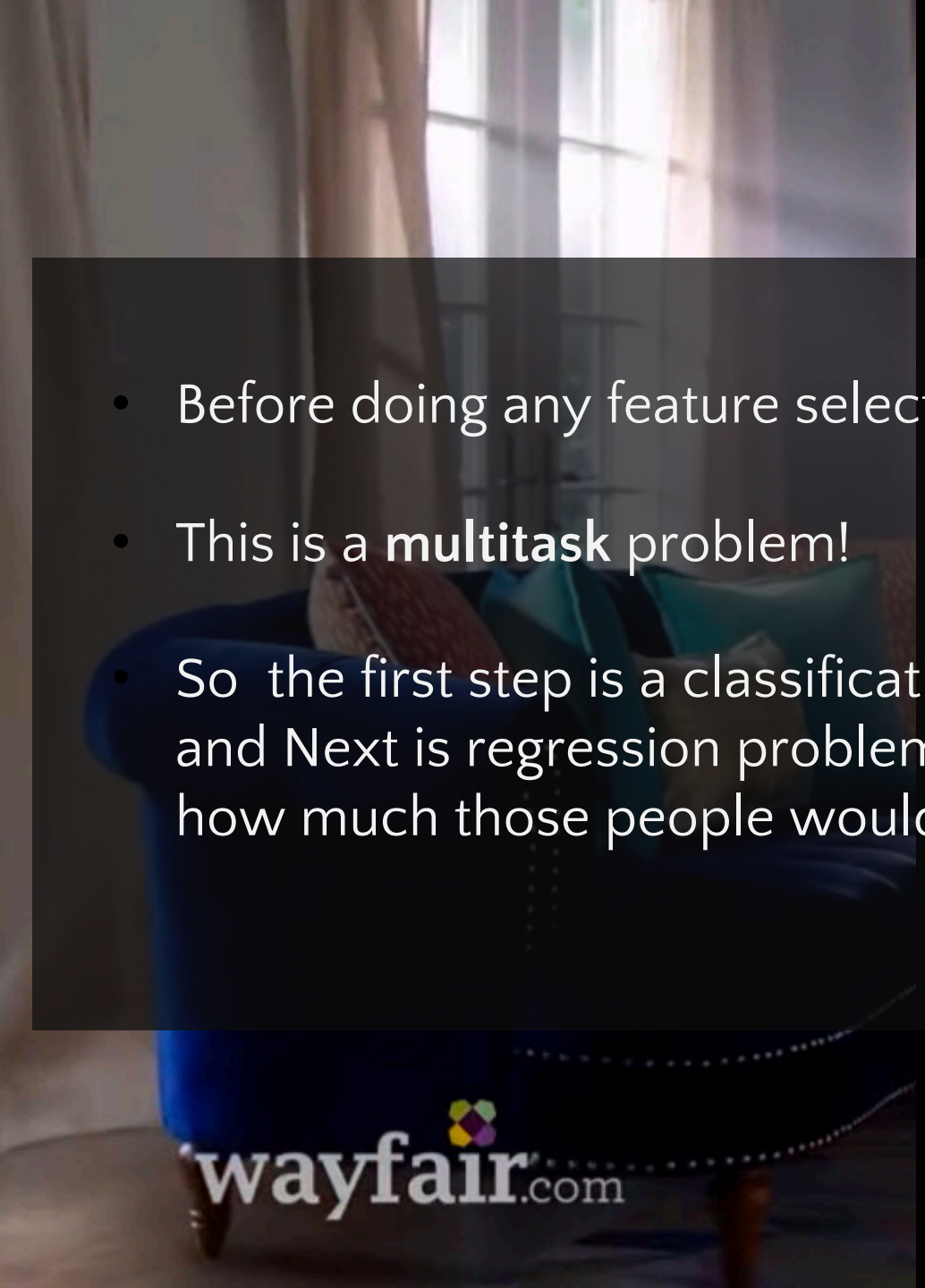
USING DATA TO BRING CUSTOMERS HOME

Predicting whether the customers coming back within 30 days and how much he/she would spend

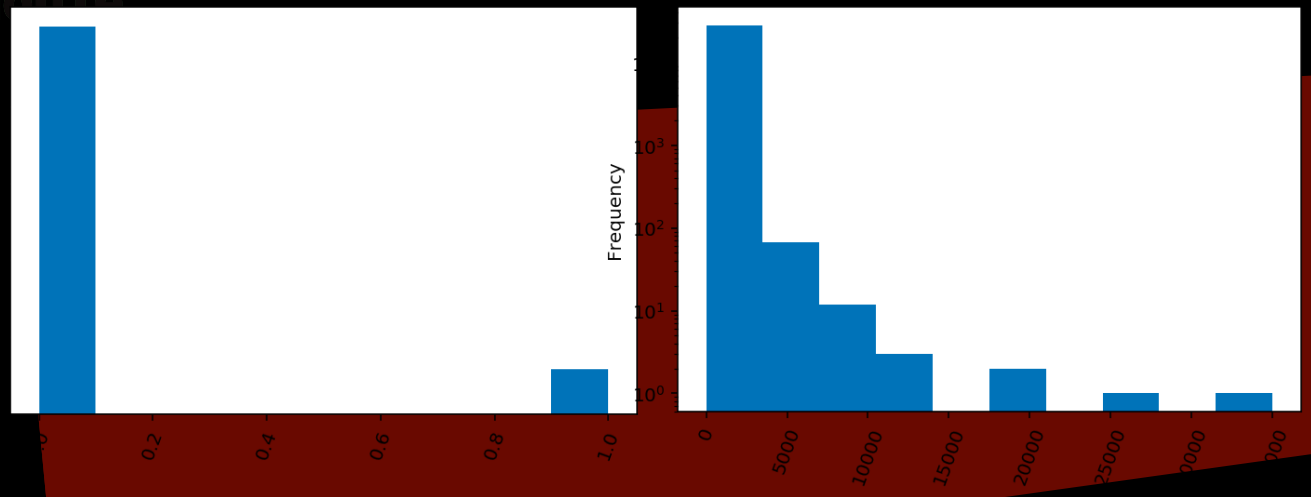
XUMING WANG
FORDHAM UNIVERSITY



- Before doing any feature selection ;
- This is a **multitask** problem!
- So the first step is a classification problem and Next is regression problem to predict how much those people would spend

Agenda

1. EDA
2. Pre-processing
3. Feature Selection
4. Modeling
5. Model evaluation



EDA

- In this dataset, target is unbalanced so we have to balanced first before modeling
- We don't have a missing data in categorical columns, so that we filing missing data with **mean value**

- In this step ;
- 1. proportion of classes
- 2. dealing with missing value

One-hot Encoding

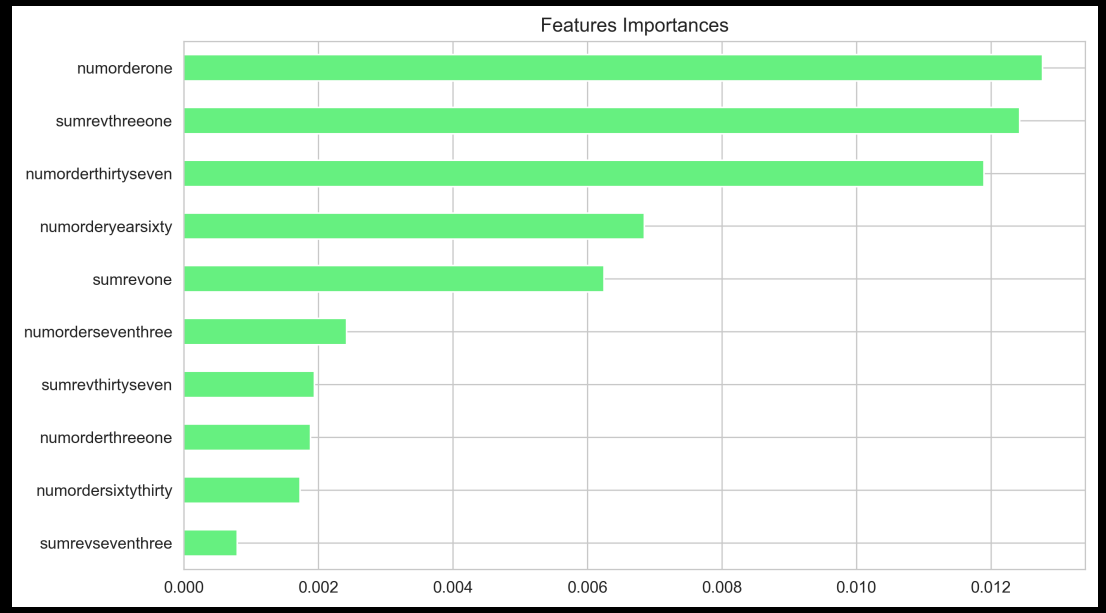
One-hot the category columns:
dummy_fields =
['roll_up',currentstatus','companytypegroup','team','customersource','accrole','num_employees', 'num_purchases_year','cost_purchases_year','enrollmentmethod']

- By using pandas getdummy!
- Keeping the data information as much as possible, coming up with dimension increasement.

Feature Selection

- Random Forest feature selection

We assume that interpretation matters, and RandomForest keeping the original columns for model/feature explain



Top 10 original features

Eventually we choose 101 features!

Tuning 5 models

- #Logitics Tuning Params

C = [0.01, 0.1, 1, 10, 100]

- #Perceptron Tuning Params

alpha = [0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0, 10, 100, 1000]

- # SVM Tunning Params
- C = [0.01, 0.1, 1, 10, 100]
- #Random Forest Tunning Params

n_estimators = [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}

- #KNN Tuning Params

n_neighbors’ :k = [1, 3, 5, 11, 21, 41, 61, 81]

Regression - Deeplearning

1. Based on the data used in classification. The Next step is using DL to fit the revenue come with the return customers.

2. Dropping the unimportant features, and building a Neural Network with Pytorch(input layer 512, hidden layers 64, output layer 1 , and the dropout as regulization)
And the loss function is MSE
3. After 300 epoch reach to a considerable result.

Save the result

THANK