Student Data Hackaton

MLewandowska

November 2020

Contents

1 Basics Look into data

1.1 Look Into Data

- read data from csv files
- group data in three tables: guests, restaurants, stars
- merge tables guests with stars into "stars-guests" table
- remove records with "?" from "stars-guests" table
- in "stars-guests" table create dummy columns for categorical variables (one-hot-encoding)
- merge tables restaurants with stars into "stars-restaurants" table
- in "stars-restaurants" table create dummy columns for categorical variables (one-hot-encoding)
- in "stars-restaurants" table create dummy columns for haurs
- check correlations (pearson coefficients) within tables
- remove columns, that are not correlated with Target (number of stars)

1.2 Data preprocessing

1. create one data set "data.csv" based on information from "Look into data".
2. create normalized data set "data-normalized.csv" where columns: weight, heitht, age, closing hour are whitened:
$$X \rightarrow \frac{X - X.mean()}{X.std()}$$
3. "data.csv" has 2936 records (rows) with 142 features (columns) each

1.3 Regression

1. read data from "data-normalized.csv" file
2. split data into training, validation and testing set
3. train linear regression model on training set and check on validation set
4. train regularized linear regression model on training set and check on validation set (grid search for best parameters)
5. train logistic regression model on training set and check on validation set
6. for logistic regression model check which features have the biggest and weights (coefficients)
7. for logistic regression model check which features have the smallest and weights (coefficients)

Conclusion:

- Linear regression and regularized linear regression are not suitable for this data
- From analysis of coefficients in logistic regression the following seems to be not important: cuisine-x-Bar, cuisine-x-Italian, cuisine-y-Pizzeria. Status "widow" is highly correlated with 0 and 1 star. Guest with Activity professional will rather give 2 stars whereas unemployment - 0 stars. Guest who has eaten cuisine Turkish will rather give 0 stars (negative coefficient for 2 stars and positive coefficient for 0 star). Restaurant Bakery is most likely to get 0 stars (negative coefficient for 2 stars and positive coefficient for 0 star)

1.4 Decision trees

1. read data from "data-normalized.csv" file
2. split data into training, validation and testing set
3. train linear decision tree classifier on training set and check on validation set
4. grid search for the best "max deepth" parameter (get max deepth =13)
5. evaluate decision tree classifier with max deepth =13, show whole tree model
6. check, which features were used for splitting

The following features have not been used for splitting:

- budget-high do nit occur
- payments other than cash
- parking-lot-valet parking
- marital-status-widow
- marital-status-single
- marital-status-married
- religion other than catholic
- cuisine-x-Bar
- cuisine-x-Italian
- open-16-18

1.5 Conclusion

Based on results from Logistic regression and decision trees, I decided to remove fro future data sets columns: martial status single, martial status married, guest cuisine Bar, guest cuisine Italian. Final data set has 2936 records (rows) with 135 features (columns) each.

2 Simple models

The following model were trained on data sets [2936 rows x 135 columns].

1. Logistic regression
2. support vector machines (with kernel poly, linear, rbf, sigmoid)
3. decision tree
4. feed forward neural network

Decision trees were trained on "data.csv", the other three models on "data-normalized.csv". In each case the best hyperparameters were found by using grid search method. The results are shown in Table:

	accuracy on training set	accuracy on training set	accuracy on training set
Logistic regression	79%	72%	72%
Decision tree	94%	89%	87%
SVN	99%	99%	95%
Neural network	97%	96%	96%

- For Support vector machines the best accuracy model achieved for kernel = 'rbf', C=10, gamma = 0.1 and coef0 = -10.
- For DecisionTreeClassifier I set parameter max-depth to 12, as further increasing max-depth didn't increase model's accuracy and cause over-fitting.
- For Neural Network model I use Feed Forward neural network model with two hidden layers: (135,40,40,3) and hiperbolic tangens activation function. The model was train using stochastic gradient descent with learning rate lr = 0.05 and momentum = 0.9.

3 Ensemble models

1. Bagging method on decision trees
2. Bagging method on decision trees, SVM and neural networks
3. XGBoost

	training set	validation set	testing set
Ensemble DT only	98%	91%	88%
Ensemble SVN / NN/ DT	98%	90%	90%
XGBoost	99%	92%	92%

- Bagging model on decision trees consists of ensemble of m=100 decision trees classifiers with max depth = 12
- Bagging model on decision trees, SVM and neural networks consists of ensemble of m=5 decision trees classifiers with max depth = 12, one SVM and one (already trained) Neural Network. SVN and Neural network aare taken with weight 5, each of decision trees classifiers have weight = 1.
- XGBClassifier is was implemented as XGBClassifier from xgboost library.

	score	name_features
f0	998	drinker_abstemious
f1	991	age
f2	914	cuisine_x_Chinese
f135	542	religion_Christian
f130	317	height
f7	309	color_green
f133	283	parking_lot_yes
f131	279	cuisine_x_Burgers
f125	251	color_white
f8	205	dress_preference_elegant
f88	175	color_black
f9	174	cuisine_y_Seafood
f22	169	open_08_10
f35	169	parking_lot_none
f19	166	cuisine_y_Mexican

(a)

	score	name_features
f0	998	drinker_abstemious
f1	991	age
f2	914	cuisine_x_Chinese
f135	542	religion_Christian
f130	317	height
f7	309	color_green
f133	283	parking_lot_yes
f131	279	cuisine_x_Burgers
f125	251	color_white
f8	205	dress_preference_elegant
f88	175	color_black
f9	174	cuisine_y_Seafood
f22	169	open_08_10
f35	169	parking_lot_none
f19	166	cuisine_y_Mexican

(b)

Figure 1: Feature importance sorted by weight (left) and gain (right) for XGBClassifier

4 Interesting data correlations

- Unemployment guests give 0 stars, students tend to give 1 or 2 stars (86%) and professionals 2 stars (in 64%)
- guests with no religion tend to give 2 stars, Jewish - 1 star and Catolics and Mormons - 1 or 2 stars.
- people with Pre-obesity ($BMI \in (25, 30)$) do not give 0 stars (in 93%),
- Very Young (born after 1991) people and old people (born before 1971) tend to give 2 stars
- Guests, who do not pay by cash tend to not give 0 stars
-