Ideatory MakeMyTrip Challenge

Submitted by

Name: Akansha Kumar

Ideatory username:akansha

Problem Definition

- Problem:- User clickstream data and information about a group of hotels is provided.
- Objective:- Segment users into a given set of classes. The classes are,
  - Backpackers
  - Family
  - Couple
- Supervised learning –
  - Set of training features and target are provided
  - A test set with features data is provided
  - Objective:- Predict the target values for the test data

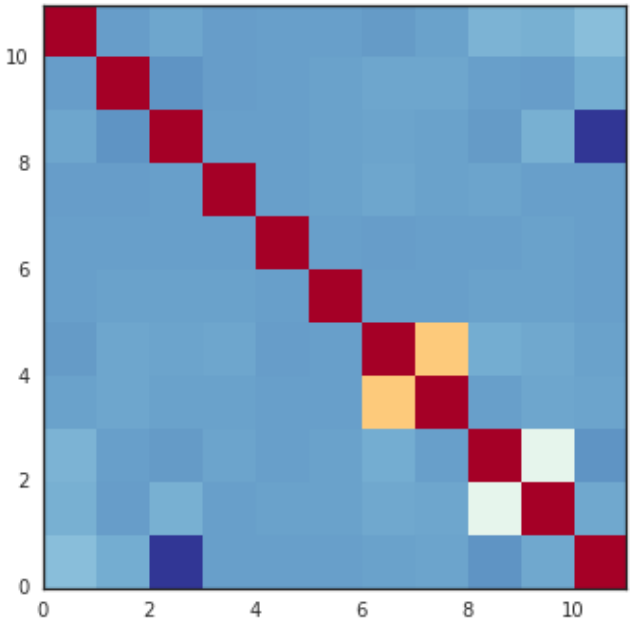
Standard Approach

Steps:-

1. Data extraction:- From csv (transfer data from csv to pandas dataframe in python)
2. Data mining:- Perform joins to combine data from multiple dataframes into a single dataframe. (join method in pandas)
3. Data manipulation:- Convert discrete non-numeric data and boolean data to integers. Develop an initial set of features
4. Break data into test data and training data for test cross-validation
5. Perform PCA and FA to determine influential features. This step decides the model features.
6. Choose classifiers –
    1. Try multiple classifiers and determine the F1 score for test cross validation
    2. For each classifier implement GridSearchCV and randomserachCV to obtain an optimum set of parameters
7. Implement the classifier on the supplied test data and generate the test target csv for submission.

Initial Features

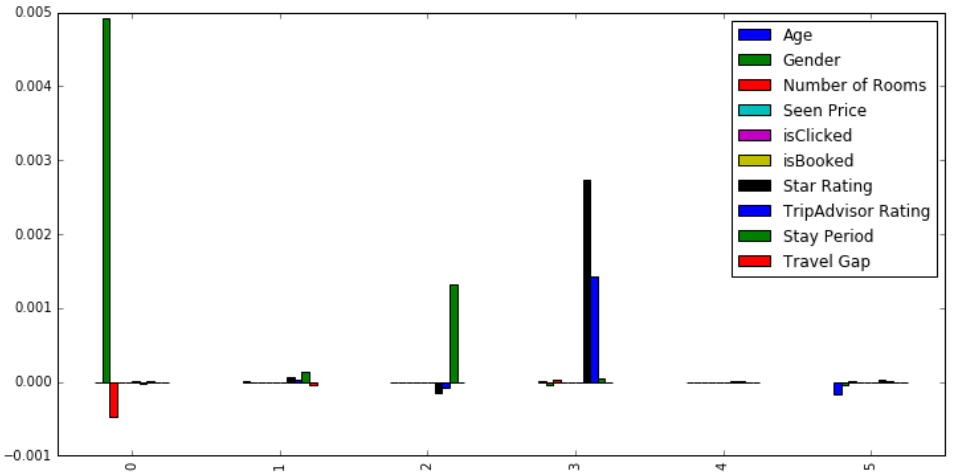
1. Age
2. Gender (Male-1, Female-0)
3. Number of Rooms
4. Seen price
5. isClicked (True-1, False-0)
6. isBooked (True-1, False-0)
7. Star Rating
8. Trip Adviser Rating
9. Stay period (Difference between Check out date and Check in date)
10. Travel Gap (Difference between booking date and Check in date)



***Correlation plot – top right corner

Feature Extraction

- Understand the influence of the features on the target
  - Correlation plot
  - Principal components (sklearn.decomposition.PCA)
  - Independent components – bottom right corner (sklearn.decomposition.FastICA)
  - Factors (sklearn.decomposition.FactorAnalysis)
- Selected Features
  - Age
  - Star Rating
  - Seen price



Cross-validation for model validation

- Prediction accuracy
- Standard approach of breaking the training data into test (test_cv) and training data (cv_train) for cross-validation.
- Use train_cv to build the regression model, and test it against test_cv.
- The score ranges from 0-1.0 where 1.0 is the best fit.
- sklearn.cross_validation_train_test_split – Split training data into random test and train data.
- Test size – 40000 (1/3rd of the total data)
- Use this validation approach to all the classifier and keep track of the scores and choose the one that has the best score.

Classifier Methods

- I have tried the following methods (all these methods are available in sklearn package in python),
  - Decision Tree Classifier
  - Gaussian Naïve Bayes
  - Support vector Machines
  - Random Forest Classifier
  - Logistic Regression
  - SGD Regressor
  - K Nearest Neighbors Classifier
  - Bernoulli Naïve Bayes
  - Linear Discriminate Analysis
  - Quadratic Discriminate Analysis
  - Ada Booster Classifier

Parameter Optimization

- All the methods described in the previous slide have hyper parameters.
- Parameters have a strong influence on the regression model hence affect the predictions.
- Cross-validation is a standard method to determine an optimum value for the hyper parameters
- Each parameter set is implemented in the model and a score is determined. Score ranges from 0 to 1.0.
- Methods:-
  - Grid based (GridSearchCV in sklearn) :- Performs a grid based on all the permutations of the parameter space (inputted by the user) and performs a search
  - Random based (RandomizedSearchCV in sklearn) :- Randomly searches for the best parameter set

Implementation of a classifier

- Standard steps:-
  1. Create an instance of the classifier
  2. Fit the training data
  3. Predict the test results

Future Work

- Predictive methods are always not complete. There is a significant scope for future work,
  - Try several other methods
  - Implement a learning based hyper parameter optimization
  - There is still room for model search
  - Search for other features not given in the data
  - More visualization
  - Implement heuristic methods for optimization

Tools

- Jupyter ipython notebook
- mysql
- python:-
  - pandas
  - sklearn
  - scipy
  - numpy
  - matplotlib