Data Analysis Project II

Aegean Airlines Passenger Satisfaction

Group No 1

B Thanushan – S14025

W. S. S. Fernando – S13990

W. M. C. B. Weerakoon – S14028

1) Abstract

Aegean Airlines wanted a model to predict whether their customers are satisfied or not with the services they are offering. Hence the main objective of this analysis is to build a statistical learning model to predict the customer satisfaction of Aegon Airlines' passengers. Also, the analysis was extended to find out the main factor that affects customer satisfaction. The dataset consisted of 129880 data of customers and the data was collected on 24 variables including the response variable. Descriptive analysis proved that the satisfaction levels of in-flight food and drink, onboard service, and seat comfort play a massive part in overall customer satisfaction. Also, it has been pointed out that passenger satisfaction is highly dependent on customer class too. Advanced analysis was carried out by fitting the models using Decision Tree Classifier, Random Forest, XGBoosting, Logistic Regression with Ridge, Lasso, and Elastic Net Penalty Techniques. The models were evaluated and compared using AUC Values and Model Accuracy Values. The model fitted using XGBoosting techniques gave the highest AUC Value and Model Accuracy which are 0.95 and 95.54% respectively. Also, it has been found out that online boarding, type of travel, in-flight wi-fi service, in-flight entertainment, on-board service, baggage handling, in-flight service, seat comfort, customer type, and check-in service are the most important features of the model with the highest accuracy.

2) Table of Contents

3) List of Figures

4) List of Tables

5) Introduction

The airline industry has evolved in its objectives over time. Earlier it was used by people only as a method to travel long distances rapidly and a method of approaching places that cannot be approached by land. But nowadays, people seek many other luxuries from an airline other than safe travel such as comfort, food, entertainment, etc. Satisfaction is not only considered as a customer's goal to be derived as a result of degrading services but also as a company's goal, as a way of getting higher customer retention rates and ways of generating profits. (Surapranata & Iskandar, 2013) Hence, with the competition in the airlines business, every airline tries to satisfy their customers on their journey in every way they can.

Measuring customer satisfaction is a key element for modern businesses as it can significantly contribute to a continuing effort of service quality improvement. In order to meet customer expectations and achieve higher quality levels, airlines need to develop a specific mechanism of passenger satisfaction measurement. (Tsafarakis, Kokotas, and Pantouvakis, 2018) Therefore, airline businesses always try to measure the levels of their customers' satisfaction in order to maintain a higher service level and obtain a high revenue in return.

6) Description of the Problem

Customer satisfaction in the service of an airline is very crucial for the market share loss of an airline. Therefore, Aegean Airlines needed to know the level of customer satisfaction of their customers and the factors that mostly affect the satisfaction of customers. This helps the company to develop strategies and plan long term to effectively handle the competition of the airline business and to increase their revenue. This study aims to build a model to predict customer satisfaction and examine the most important factors that influence the passenger's expectation from an airline service from the given dataset. Finally, creating a website for Aegean airlines that will predict whether a customer is satisfied or not through the built model.

7) Description of the Dataset

The dataset is contained airline passenger satisfaction survey data. It consists of 129880 customer feedbacks containing 24 features including both qualitative and quantitative. Satisfaction is the response variable in this study.

No.	Variable name	Description	Type of the variable
1.	satisfaction	Airline satisfaction level (Satisfaction, neutral, or dissatisfaction)	qualitative
2.	id	Customer identification number	
3.	gender	Gender of the passengers (Female, Male)	qualitative
4.	age	The actual age of the passengers	quantitative
5.	type_of_travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)	qualitative
6.	customer_class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)	qualitative
7.	customer_Type	The customer type (Loyal customer, disloyal customer)	qualitative
8.	flight_distance	The flight distance of this journey	quantitative
9.	inflight_wifi_service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)	qualitative
10.	ease_of_online_booking	Satisfaction level of online booking	qualitative
11.	inflight_service	Satisfaction level of inflight service	qualitative
12.	online_boarding	Satisfaction level of online boarding	qualitative
13.	inflight_entertainment	Satisfaction level of inflight entertainment	qualitative
14.	food_and_drink	Satisfaction level of Food and drink	qualitative
15.	seat_comfort	Satisfaction level of Seat comfort	qualitative
16.	leg_room_service	Satisfaction level of Leg room service	qualitative
17.	baggage_handling	Satisfaction level of baggage handling	qualitative
18.	gate_location	Satisfaction level of Gate location	qualitative
19.	cleanliness	Satisfaction level of Cleanliness	qualitative
20.	check_in_service	Satisfaction level of Check-in service	qualitative
21.	departure_delay_in_minutes	Minutes delayed when departure	quantitative
22.	arrival_delay_in_minutes	Minutes delayed when Arrival	quantitative
23.	onboard_service	Satisfaction level of onboard service	qualitative
24.	departure_arrival_time_convenient	Satisfaction level of departure and arrival time convenient	qualitative

Table 7-1: Description of the Dataset

8) Important Results of the Descriptive Analysis

According to the Onboard Service vs. Satisfaction graph, we can see that most of the customers who selected onboard service satisfaction levels 4 and 5 are satisfied with the airline. Also, most of the customers who selected onboard service satisfaction levels 1,2, and 3 are neutral or dissatisfied with the airline. Thus, it is important to consider onboard service satisfaction to improve airline passenger

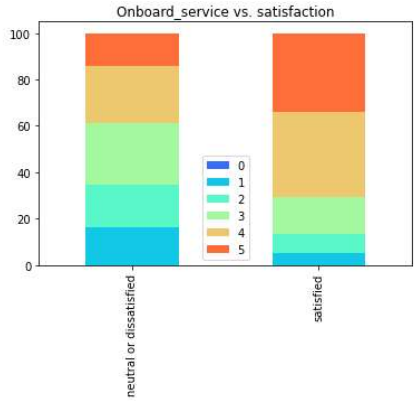

Figure 8-2: Stacked Bar Graph of Onboard Service vs. Satisfaction

satisfaction. Seat comfort is defined as the degree of passenger satisfaction in terms of comfort while sitting in an airliner. The measurements of seat comfort are based on legroom, seat recline, seat width, aisle space, and ease of video viewing. Out of the three factors that determine customer satisfaction in the airline industry, seat comfort is the most significant factor that affects customer satisfaction, following by food and beverages and staff services respectively. (Zaharias, 2016) Here we can see that most of the customers selected 3 to 5 levels. Hence, we can conclude that the passengers are satisfied with the seat comfort of the fights. Most customers who have selected seat comfort satisfaction levels as 4 and 5 are also satisfied with the airline. Also, most customers who have selected seat comfort satisfaction levels as 1,2, and 3 are neutral or dissatisfied with the airline. So, we can say that there is a relationship between passenger satisfaction on the airline and seat comfort of the flights. Here we can see that most of the customers who traveled long flight distances are satisfied with the airline. The reason for that is people look for good onboard service and seat comfort when they are traveling long distances. Since most of the customers are satisfied with onboard service and seat comfort, they tend to satisfy with the airline.

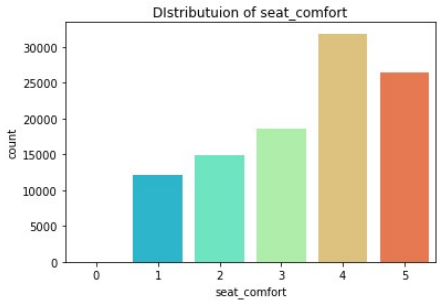

Figure 8-1: Bar Graph of Seat Comfort

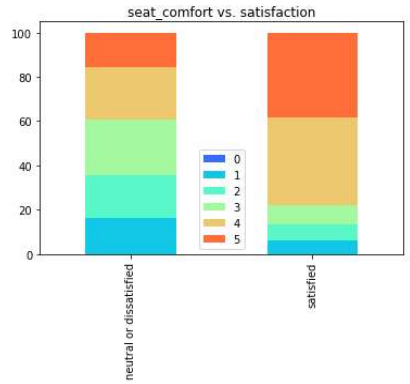

Figure 8-3: Stacked Bar Graph of Seat Comfort vs. Satisfaction

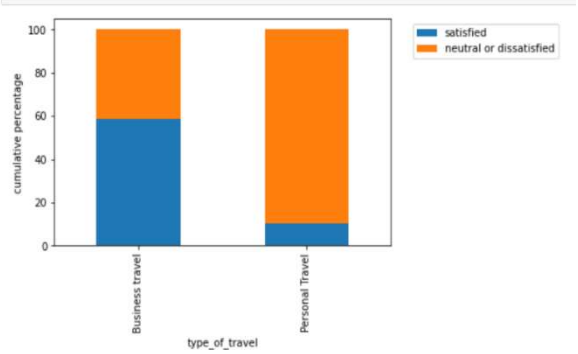

Figure 8-4: Stacked Bar Graph of Type of Travel vs. Satisfaction

The business class has upgraded services than the economy class and hence the customers who travel in the business class are most probably satisfied with the airline services. (Travel Tips - USA Today, 2021) (Covington Travel, 2021) This has been proven from our study too.

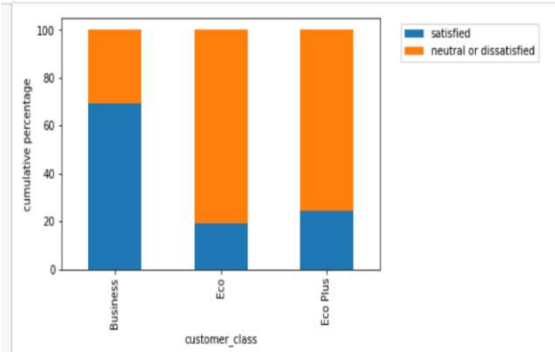

Figure 8-5: Stacked Bar Graph of Customer Class vs. Satisfaction

We can see that the people who travel in business class are more satisfied with the services than the people who travel in economy and economy plus class. We can also see that the people travel for personal reason are more dissatisfied with the airlines. The reason for that is the people who travel personal purpose mostly travel in Economy or Economy plus class and people travel for business purposes travel in business class. Therefore, it can be concluded that the business class services better than the other two.

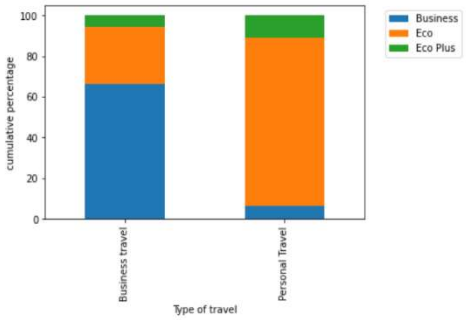

Figure 8-7: Stacked Bar Graph of Type of Travel vs. Satisfaction

The airline frequent flyer program is a customer loyalty program in which the airline provides offers and benefits, such as free flights and products, based on the number of miles a customer has flown. Customers choose to fly on airlines where they have accumulated the most miles. At the same time, loyalty is not exclusive since customers can enroll in an unlimited amount of customer loyalty programs. Customer loyalty is rooted in economic benefits to the customer, which drives repeat purchases. Passenger satisfaction had a significant effect on passenger loyalty. (Namukasa, 2013) As earlier found out, the people who fly in economy class are more dissatisfied than the people who fly in business class. From the graph, it is clear that most disloyal customers travel in economy class. This can be a reason for the majority dissatisfaction among the disloyal customers. Thus, we have to give more concern to develop economy class facilities in order to maximize customer loyalty and satisfaction.

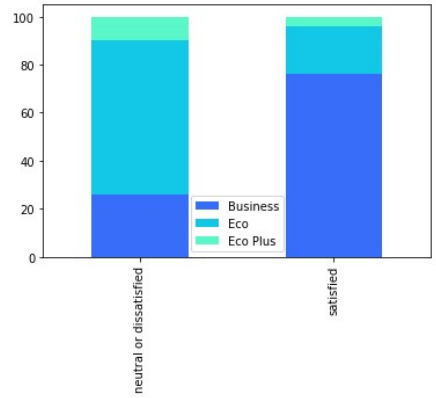

Figure 8-6: Stacked Bar Graph of Satisfaction vs. Customer Class

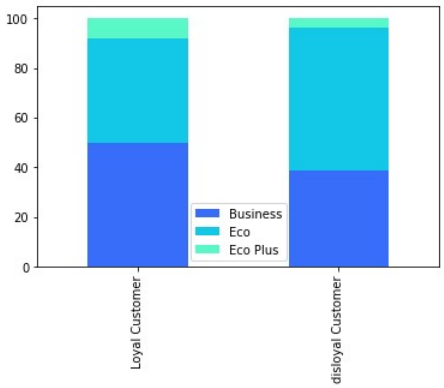

Figure 8-8: Stacked Bar Graph of Customer type vs. Customer Class

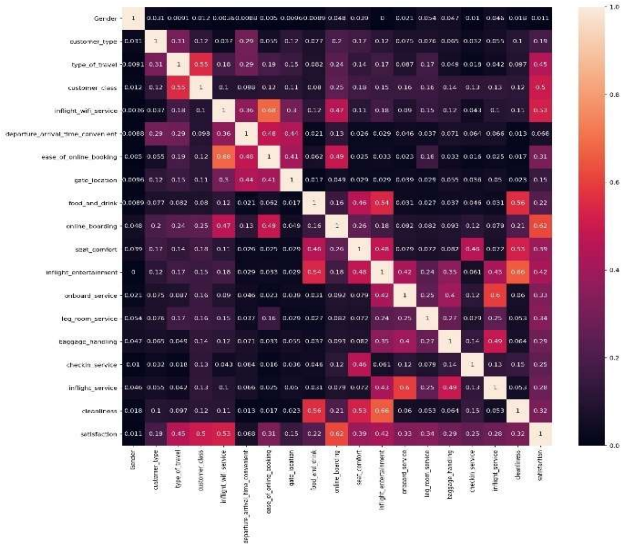

Figure 8-10: Correlation Plot of Categorical Variables

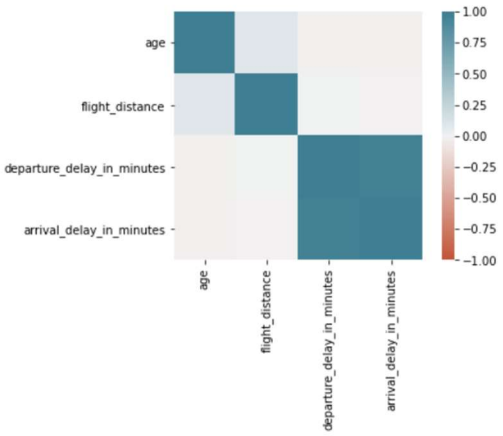

Figure 8-9: Correlation Plot of Continuous Variables

Clearly, there are some correlations between variables. They are ease of online booking and inflight Wi-Fi service, Online boarding and satisfaction, cleanliness and inflight entertainment, onboard service, and inflight service. Also, there is a correlation between Arrival Delay in Minutes and Departure Delay in Minutes.

9) Important Results of the Advanced Analysis

The main objective of the advanced analysis was to find the most accurate model to predict whether the customers are satisfied with the airlines or not. Also, the advanced analysis is focused on finding the factors which are most influential in customers to be satisfied with the airline. Logistic Regression with Ridge, Lasso, and Elastic Net Penalty Methods, Decision Tree Classifier, Random Forest, and XGBoosting were used in the advanced analysis. The AUC value and the Model Accuracy Score of models are used in evaluating the models.

The AUC value of the Decision Tree Classifier is 0.93 with 93.83% model accuracy. We have found out that the type of travel, in-flight wi-fi

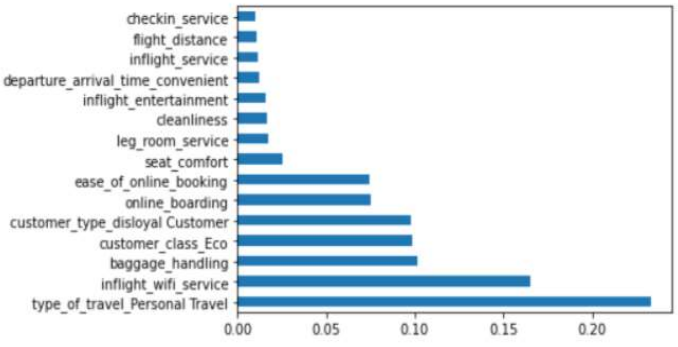

Figure 9-1: Feature Importance Plot of Decision Tree Model

service, baggage handling, customer class, customer type, online boarding, and ease of online booking are the important features of the model.

Random Forest Classifier has an AUC value of 0.94 and a model accuracy of 94.43%. It has been also found out that online boarding, in-flight wi-fi service, type of travel, customer class, in-flight entertainment, seat comfort, customer type, ease of online booking, legroom service, onboard service, cleanliness, and flight distance are the important features of the model.

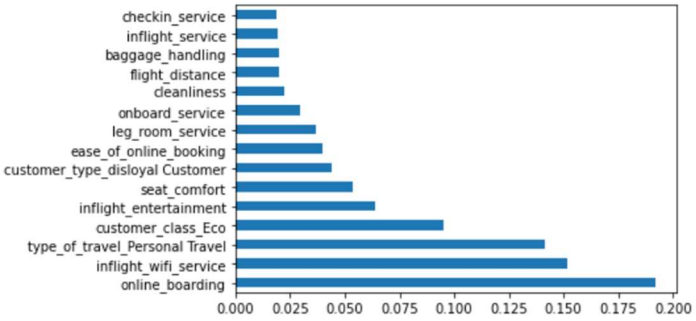

Figure 9-2: Feature Importance Plot of Random Forest Model

The AUC value and the model accuracy of XGBoosting are 0.95 and 95.54% respectively. Online boarding, type of travel, in-flight wi-fi service, in-flight entertainment, onboard service, baggage handling, in-flight service, seat comfort, customer type, and check-in service are identified as the important features of the model.

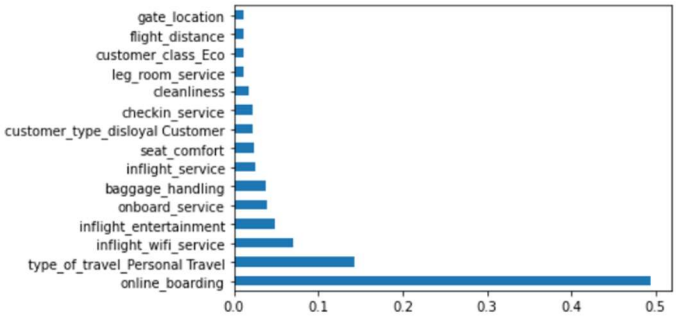

Figure 9-3: Feature Importance Plot of XGBoosting Model

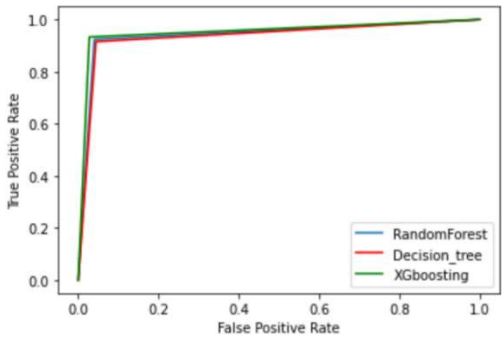

Figure 9-4: ROC Curves

Models are fitted with applying Ridge, Lasso and Elastic Net penalties for the Logistic Regression and they had AUC values of 0.87, 0.71, 0.72 respectively, and 87%, 72%, 72.22% model accuracies respectively.

MODEL	AUC Value	Accuracy
Decision Tree	0.93	93.83%
Random Forest	0.94	94.43%
XGBoosting	0.95	95.54%
Logistic Regression with Ridge Penalty	0.87	87%
Logistic Regression with Lasso Penalty	0.71	72%
Logistic Regression with Elastic Net Penalty	0.72	72.22%

Table 9-1: AUC Value and Accuracy of the Models

Since The model fitted using the XGBoosting technique has the highest AUC value and the highest model accuracy out of all models, we choose that as the most accurate model.

```
[[14324   405]
 [  753 10494]]
              precision    recall  f1-score   support

           0       0.95      0.97      0.96     14729
           1       0.96      0.93      0.95     11247

    accuracy                           0.96     25976
   macro avg       0.96      0.95      0.95     25976
weighted avg       0.96      0.96      0.96     25976
```

Figure 9-5: Confusion Matrix and Classification Report of XGBoosting Model

10) Issues Encountered and Proposed Solutions

In our dataset, we had comparatively few missing values for the variable 'arrival delay in minutes' which is skewed. It has been recommended to use the median for skewed data with outliers. (Data Analytics) Therefore, we have imputed the missing values using the median. We had a significant association between predictor variables. It has been stated to use regularization like Ridge, Lasso, and Elastic Net for Logistic Regression when there are associations between variables. (Datacamp) Therefore, we have fitted logistic regression with ridge penalty, lasso penalty, and elastic net. Tree-based classifiers are less sensitive to multi-collinearity and do not affect much in prediction. (Wiley Online Library) Hence, we have fitted Decision Tree, Random Forest, and XGBoosting.

11) Discussion and Conclusions

The main objective of the analysis was to predict the customer satisfaction of the passengers of Aegon Airlines. Another objective of the study was to identify which factors affect the satisfaction of the customers the most.

There were missing values in the variable 'arrival delay in minutes', and those values were imputed using the median of the present data of the variable.

The model fitted using the XGBoosting technique gave the highest AUC Value and the Model Accuracy. The important features identified from the model are Online boarding, type of travel and in-flight wi-fi service, in-flight entertainment, onboard service, baggage handling, in-flight service, seat comfort, customer type, and check-in service.

12) References

Tsafarakis, S., Kokotas, T. and Pantouvakis, A., 2018. A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. *Journal of Air Transport Management*, 68, pp.61-75.

Zaharias, B. (2016). ANALYSING CUSTOMER SATISFACTION IN THE AIRLINE INDUSTRY.

Han, H. and Hwang, J., 2014. In-flight physical surroundings: quality, satisfaction, and traveller loyalty in the emerging low-cost flight market. *Current Issues in Tourism*, 20(13), pp.1336-1354.

Mohd Zahari, M., Salleh, N., Kamaruddin, M. and Kutut, M., 2011. In-flight Meals, Passengers' Level of Satisfaction and Re-flying Intention.

Travel Tips - USA Today. 2021. *How to Travel on Business Class*. [online] Available at: <https://traveltips.usatoday.com/travel-business-class-35337.html>.

Covington Travel. 2021. *5 Reasons You Need to Fly Business Class - Covington Travel*. [online] Available at: <https://www.covingtontravel.com/2019/02/5-reasons-need-fly-business-class/>.

Feature Importance in Logistic Regression for Machine Learning Interpretability - Sefik Ilkin Serengil. (2021)., from https://sefiks.com/2021/01/06/feature-importance-in-logistic-regression/

How to tune a Decision Tree?. (2021)., from https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

Trees. (2021). Cost Complexity Pruning in Decision Trees | Decision Tree., from https://www.analyticsvidhya.com/blog/2020/10/cost-complexity-pruning-decision-trees/

InDepth: Parameter tuning for Decision Tree. (2021)., from https://medium.com/@mohtedibf/indepth-parameter-tuning-for-decision-tree-6753118a03c3

13) Appendix and R Code

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.metrics import roc curve, auc
from sklearn.metrics import accuracy score
# In[23]:
clf = DecisionTreeClassifier(criterion='gini',
max_depth=8, random_state=0)
# In[24]:
clf.fit(X_train,Y_train)
# In[25]:
y_pred= clf.predict(X_test)
# In[26]:
print('Model accuracy score: {0:0.4f}'.
format(accuracy score(Y test, y pred)))
# In[27]:
y train pred = clf.predict(X train)
# In[28]:
print('Model accuracy score: {0:0.4f}'.
format(accuracy_score(Y_train, y_train_pred)))
# ### Using GridSearch to tune the paramaters of
Decision tree
# In[29]:
from sklearn.model selection import GridSearchCV
# In[30]:
params = {
    'max depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [5, 10, 20, 50, 100],
    'max_features' : ['auto', 'sqrt', 'log2']
}
# In[31]:
clf_TUNED = DecisionTreeClassifier(random_state=20)
# In[32]:
grid search = GridSearchCV(estimator=clf TUNED,
                           param_grid=params,
                                  cv=5, n_jobs=-1, verbose=1,
scoring = "accuracy")

# In[33]:
grid search.fit(X train, Y train)
# In[34]:
grid_search.best_params_
# In[35]:
grid_search.best_score_
# ##### Use the best tuned paramater and fit the model
again
# In[36]:
DT clf = DecisionTreeClassifier(max depth = 20,
max features = 'auto', min samples leaf = 10,
random state=20)
# In[37]:
DT_clf.fit(X_train, Y_train)
# In[38]:
Y_pred = DT_clf.predict(X_test)
# In[81]:
print('Model accuracy score for decision tree:
{0:0.4f}'. format(accuracy_score(Y_test, Y_pred)))
# In[40]:
Pred trainX = DT clf.predict(X train)
# In[41]:
print('Model accuracy score: {0:0.4f}'.
format(accuracy_score(Y_train, Pred_trainX)))
# In[42]:
from sklearn.metrics import roc_curve, auc
false_positive_rate, true_positive_rate, thresholds =
roc_curve(Y_test, Y_pred)
roc_auc = auc(false_positive_rate, true_positive_rate)
roc auc
# In[43]:
# plot the roc curve for the model
plt.plot(false_positive_rate, true_positive_rate,
label='Decision tree')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
# In[44]:
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification report
# In[45]:
print(classification report(Y test, Y pred))
# In[46]:
classification_report(Y_test, Y_pred)
# In[47]:
confusion_matrix(Y_test, Y_pred)
# In[70]:
print(confusion_matrix(Y_test, Y_pred))
print(classification_report(Y_test, Y_pred))
# In[48]:
features = X train.columns
importances = DT clf.feature importances
indices = np.argsort(importances)
plt.title('Feature Importances')
plt.barh(range(len(indices)), importances[indices],
color='b', align='center')
plt.yticks(range(len(indices)), [features[i] for i in
indices])
plt.xlabel('Relative Importance')
plt.show()
# In[49]:
```

```
feat_importances =
pd.Series(DT_clf.feature_importances_,
index=X_train.columns)
feat importances.nlargest(15).plot(kind='barh')
#https://stackoverflow.com/questions/44101458/random-
forest-feature-importance-chart-using-python
# https://github.com/WillKoehrsen/Machine-Learning-
Projects/blob/master/Random%20Forest%20Tutorial.ipynb
<br/>
# # Random FOrest
# In[50]:
from sklearn.ensemble import RandomForestClassifier
# In[51]:
RF =
RandomForestClassifier(n estimators=150,max depth=10,
random state=20 ,verbose = 1)
# In[52]:
RF.fit(X_train,Y_train)
# In[53]:
y_pred_RF = RF.predict(X_test)
print('Model accuracy score: {0:0.4f}'.
format(accuracy_score(Y_test,y_pred_RF)))
# In[54]:
feat imp RF = pd.Series(RF.feature importances ,
index=X train.columns)
feat imp RF.nlargest(15).plot(kind='barh')
# In[80]:
feat_imp_RF.nlargest(12)
# In[56]:
important_feat = list(feat_imp_RF.nlargest(12).index)
important_feat.append('customer_class_Eco Plus')
important_feat
# In[57]:
X train imp = X train.loc[:,important feat]
X test imp = X test.loc[:,important feat]
# In[58]:
X train.info()
# In[59]:
param_grid_RF = {
    'n_estimators': [100,200,300,400,500],
    'max_features': ['auto', 'sqrt', 'log2'],
    'max_depth' : [6,7,8,9,10],
    'criterion' :['gini', 'entropy']
}
# In[60]:
RFC = RandomForestClassifier(random state=20)
# In[62]:
Grid_rf= GridSearchCV(estimator=RFC,
param_grid=param_grid_RF, cv= 5,n_jobs=-1,verbose
=1,scoring='roc_auc')
Grid_rf.fit(X_train_imp, Y_train)
# In[63]:
Grid_rf.best_params_
# In[64]:
Grid rf.best estimator
# In[65]:
Grid rf.best score
# In[66]:
Best_RF = RandomForestClassifier(max_depth=10,
n_estimators=400, random_state=20,criterion = 'gini',
max_features='auto')
# In[67]:
Best RF.fit(X train imp, Y train)
# In[82]:
Y pred BRF = Best RF.predict(X test imp)
print('Model accuracy score for Random forest:
{0:0.4f}'. format(accuracy_score(Y_test,Y_pred_BRF)))
# In[69]:
print(confusion_matrix(Y_test,Y_pred_BRF))
print(classification_report(Y_test,Y_pred_BRF))
# In[72]:
#false positive rate and true positive rate
fpr rf, tpr rf, thresholds rf =
roc curve(Y test,Y pred BRF)
roc auc = auc(fpr rf, tpr rf)
roc auc
# In[83]:
# plot the roc curve for the  random forest model
plt.plot(fpr_rf, tpr_rf, label='RandomForest')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
# In[78]:
# plot the roc curve for the  random forest model
plt.plot(fpr_rf, tpr_rf, label='RandomForest')
plt.plot(false positive rate, true positive rate,
label='Decision tree',color='red')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
# # Xgboosting
# #### references
# https://www.mikulskibartosz.name/xgboost-
hyperparameter-tuning-in-python-using-grid-search/
<br/>
# https://www.analyticsvidhya.com/blog/2018/09/an-end-
to-end-guide-to-understand-the-math-behind-xgboost/
<br/>
#
https://www.analyticsvidhya.com/blog/2016/03/complete-
guide-parameter-tuning-xgboost-with-codes-python/ <br/>
# https://www.datacamp.com/community/tutorials/xgboost-
in-python <br/>
# In[85]:
import xgboost as xgb
from xgboost import XGBClassifier
# In[86]:
xgb_model = XGBClassifier(
random_state=20,learning_rate=0.01,eval_metric='auc')
# fit the model with the training data
xgb model.fit(X train,Y train)
# In[87]:
feat imp xgb =
pd.Series(xgb model.feature importances ,
index=X train.columns)
feat imp xgb.nlargest(15).plot(kind='barh')
# In[88]:
feat_imp_xgb.nlargest(10)
# In[89]:
imp_feat_xgb = list(feat_imp_xgb.nlargest(10).index)
imp feat xgb
# In[90]:
X train xgb = X train.loc[:,imp feat xgb]
X test xgb = X test.loc[:,imp feat xgb]
# [22:38:04] WARNING:
C:/Users/Administrator/workspace/xgboost-
win64_release_1.4.0/src/learner.cc:1095: Starting in
XGBoost 1.3.0, the default evaluation metric used with
the objective 'binary:logistic' was changed from
'error' to 'logloss'. Explicitly set eval_metric if
you'd like to restore the old behavior.
# GridSearchCV(cv=5,
#              estimator=XGBClassifier(base score=None,
booster=None,
#
colsample_bylevel=None,
#
colsample_bynode=None,
#
colsample_bytree=None, gamma=None,
#                                     gpu id=None,
importance type='gain',
#
interaction constraints=None,
#
learning_rate=None, max_delta_step=None,
#                                     max_depth=None,
min_child_weight=None,
#                                     missing=nan,
monotone_constraints=None,
#
n estimators=100, n jobs=None,
#
num parallel tree=None, random state=20,
#                                     reg_alpha=None,
reg_lambda=None,
#
scale_pos_weight=None, subsample=None,
#
tree_method=None, validate_parameters=None,
#                                     verbosity=None),
#              n jobs=-1,
#              param grid={'gamma': [0.0, 0.1, 0.2],
#                          'learning rate': [0.1, 0.01,
0.03, 0.05],
#                          'max_depth': [6, 7, 8, 9,
10],
#                          'n_estimators': [100, 200,
300, 400, 500],
#                          'subsample': [0.6, 0.7]},
#              verbose=1)
# In[74]:
grid search xgb.best params
# In[75]:
grid_search_xgb.best_estimator_
# In[91]:
final_xgb = XGBClassifier(base_score=0.5,
booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1,
gamma=0.2, gpu id=-1,
              importance type='gain',
interaction constraints='',
              learning rate=0.03, max delta step=0,
max_depth=9,
              min_child_weight=1,
monotone_constraints='()',
              n_estimators=400, n_jobs=8,
num_parallel_tree=1, random_state=20,
              reg alpha=0, reg lambda=1,
scale pos weight=1, subsample=0.6,
              tree method='exact',
validate parameters=1, verbosity=None)
# In[92]:
final_xgb.fit(X_train_xgb, Y_train)
# In[93]:
Y_pred_XGB = final_xgb.predict(X_test_xgb)
print('Model accuracy score for XG_boosting: {0:0.4f}'.
format(accuracy score(Y test,Y pred XGB)))
# In[94]:
print(confusion matrix(Y test,Y pred XGB))
print(classification report(Y test,Y pred XGB))
# In[95]:
#false positive rate and true positive rate
fpr_xgb, tpr_xgb, thresholds_xgb =
roc_curve(Y_test,Y_pred_XGB)
roc_auc_xgb = auc(fpr_xgb, tpr_xgb)
roc auc xgb
# In[96]:
# plot the roc curve for the  random forest model
plt.plot(fpr xgb, tpr xgb, label='XGboosting')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
# In[97]:
# plot the roc curve for the  random forest model
plt.plot(fpr rf, tpr rf, label='RandomForest')
plt.plot(false positive rate, true positive rate,
label='Decision tree',color='red')
plt.plot(fpr xgb, tpr xgb,
label='XGboosting',color='green')
# axis labels
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
# show the legend
plt.legend()
# show the plot
plt.show()
from sklearn.linear_model import RidgeClassifier
from sklearn.model_selection import cross_val_score
from sklearn import metrics
clf = RidgeClassifier().fit(X_train, Y_train)
clf.score(X_train, Y_train)


Ypred = clf.predict(X_test)
clf.score(X_test,Ypred)
auc = metrics.roc_auc_score(Y_test,Ypred)
auc
clf.get_params()
clf.coef_
from sklearn.linear_model import LogisticRegression
model_elastic = LogisticRegression(penalty =
'elasticnet', solver = 'saga', l1_ratio = 0.5)
model_elastic.fit(X_train, Y_train)
feat_labels = X_train.columns
feature_importance=pd.DataFrame({'feat_labels':list(X_t
rain.columns),'feature_importance':[abs(i) for i in
model_elastic.coef_[0]]})
feature_importance.sort_values('feature_importance',asc
ending=False)
ax =
feature_importance.sort_values('feature_importance',asc
ending=False).plot.barh(x='feat_labels',
y='feature_importance', figsize = (18,6))
plt.show()
Y_predict = model_elastic.fit(X_train,
Y_train).predict(X_test)
from sklearn.metrics import classification_report,
confusion_matrix
print(confusion_matrix(Y_test, Y_predict))
print(classification_report(Y_test, Y_predict))
from sklearn.metrics import roc_curve, roc_auc_score,
accuracy_score
test_accuracy = accuracy_score(Y_test, Y_predict)*100
test_auc_roc = roc_auc_score(Y_test, Y_predict)*100
print('Test accuracy: %.4f %%' % test_accuracy)
```