Using the data in the link below, attempt to model a customer’s propensity to join our loyalty program

Ayan Karim

In [1]:
```
# Import Dependencies
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.feature_selection import RFE
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier, GradientBoostingCla
ssifier
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import GaussianNB
from sklearn import metrics
from sklearn.metrics import confusion_matrix

from scipy.stats import boxcox
```

In [2]:
```
df = pd.read_csv('customers_data.csv')
```

Explore Data

In [3]:
```
df.head()
```

Out[3]:

	Unnamed: 0	purch_amt	gender	card_on_file	age	days_since_last_purch	loyalty
0	0	19.58	male	no	31.0	35.0	False
1	1	65.16	male	yes	23.0	61.0	False
2	2	40.60	female	no	36.0	49.0	False
3	3	38.01	male	yes	47.0	57.0	False
4	4	22.32	female	yes	5.0	39.0	False

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 120000 entries, 0 to 119999
Data columns (total 7 columns):
Unnamed: 0                 120000 non-null int64
purch_amt                  120000 non-null float64
gender                     120000 non-null object
card_on_file               120000 non-null object
age                        120000 non-null float64
days_since_last_purch      120000 non-null float64
loyalty                    120000 non-null bool
dtypes: bool(1), float64(3), int64(1), object(2)
memory usage: 5.6+ MB
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Unnamed: 0', 'purch_amt', 'gender', 'card_on_file', 'age',
       'days_since_last_purch', 'loyalty'],
      dtype='object')
```

In [6]:
```
# Drop Unnamed column
df = df.drop(['Unnamed: 0'], axis=1)
```

In [7]:
```
df.describe()
```

Out[7]:

	purch_amt	age	days_since_last_purch
count	120000.000000	120000.000000	120000.000000
mean	44.036234	25.803008	56.605908
std	20.473148	10.153072	16.422187
min	-43.950000	-22.000000	-9.000000
25%	30.210000	19.000000	45.000000
50%	43.970000	26.000000	57.000000
75%	57.830000	33.000000	68.000000
max	142.200000	71.000000	125.000000

Clean Data

In [8]:
```
# Remove unrealistice values for continuous variables
df.purch_amt = df.purch_amt[df.purch_amt > df.purch_amt.quantile(.25)]
df.age = df.age[df.age > df.age.quantile(.25)]
df.days_since_last_purch = df.days_since_last_purch[df.days_since_last_p
urch > df.days_since_last_purch.quantile(.25)]

# Remove rows with null values
df = df.dropna()
```

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 47455 entries, 1 to 119999
Data columns (total 6 columns):
purch_amt                47455 non-null float64
gender                   47455 non-null object
card_on_file             47455 non-null object
age                      47455 non-null float64
days_since_last_purch    47455 non-null float64
loyalty                  47455 non-null bool
dtypes: bool(1), float64(3), object(2)
memory usage: 2.2+ MB
```

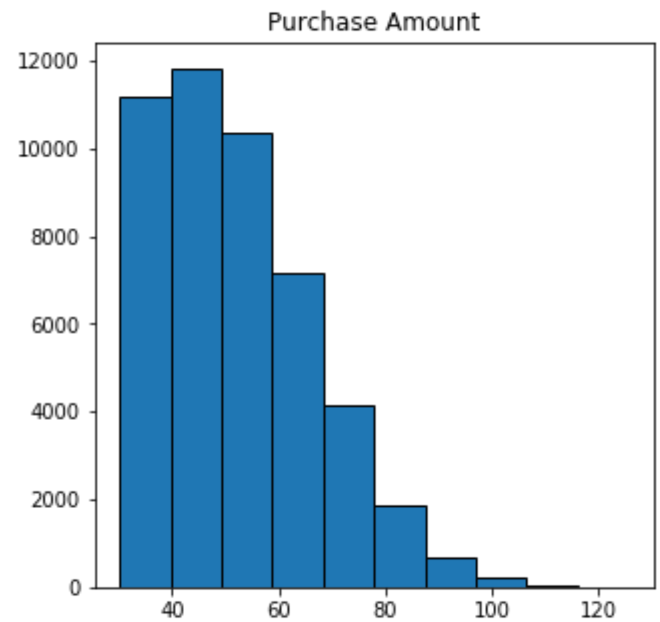
In [10]:
```
df.describe()
```

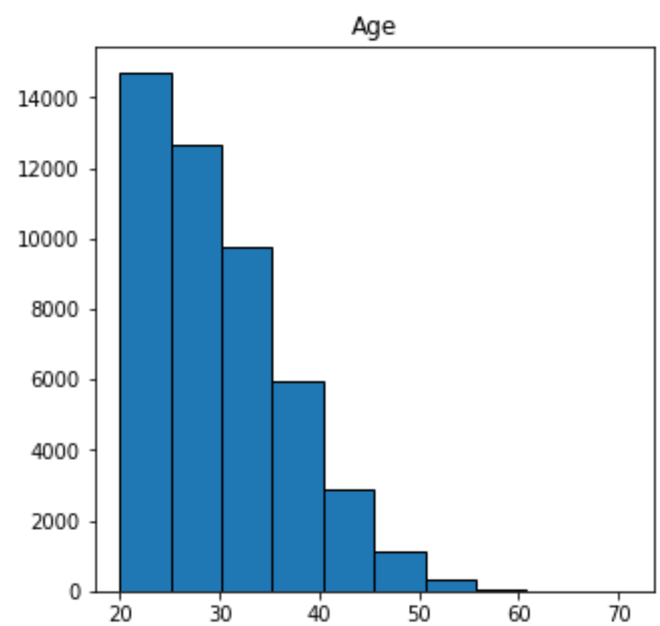
Out[10]:

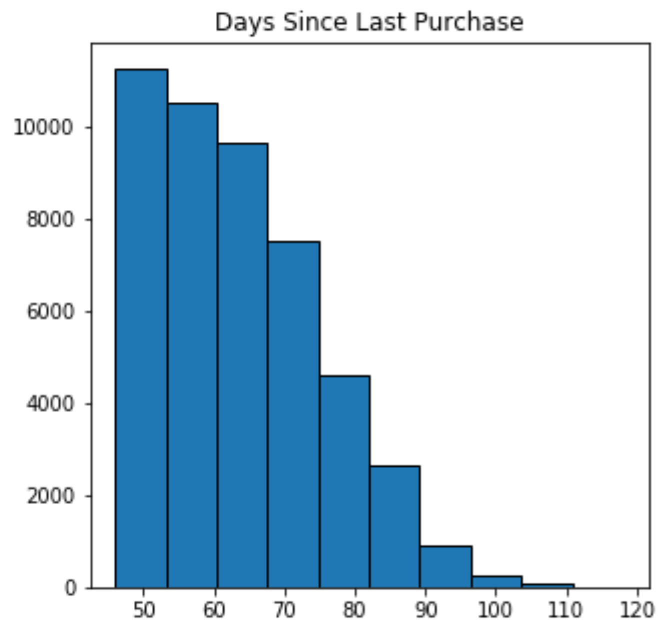
	purch_amt	age	days_since_last_purch
count	47455.000000	47455.000000	47455.00000
mean	52.051808	30.093373	63.38940
std	14.710210	7.180507	11.81037
min	30.220000	20.000000	46.00000
25%	40.300000	24.000000	54.00000
50%	49.870000	29.000000	62.00000
75%	61.360000	35.000000	71.00000
max	125.530000	71.000000	118.00000

Visualize Data

In [11]:
```
# Visualize distribution of continuous variables
plt.figure(figsize=(5,5))
plt.hist(df.purch_amt, edgecolor = 'k')
plt.title('Purchase Amount')
plt.show()

plt.figure(figsize=(5,5))
plt.hist(df.age, edgecolor = 'k')
plt.title('Age')
plt.show()

plt.figure(figsize=(5,5))
plt.hist(df.days_since_last_purch, edgecolor = 'k')
plt.title('Days Since Last Purchase')
plt.show()
```


Purchase Amount
12000
10000
8000
6000
4000
2000
0
40
60
80
100
120


Age
14000
12000
10000
8000
6000
4000
2000
0
20
30
40
50
60
70



In [12]:
```
df.head()
```

Out[12]:

	purch_amt	gender	card_on_file	age	days_since_last_purch	loyalty
1	65.16	male	yes	23.0	61.0	False
2	40.60	female	no	36.0	49.0	False
3	38.01	male	yes	47.0	57.0	False
6	43.96	male	yes	36.0	64.0	False
9	93.63	female	no	40.0	47.0	True

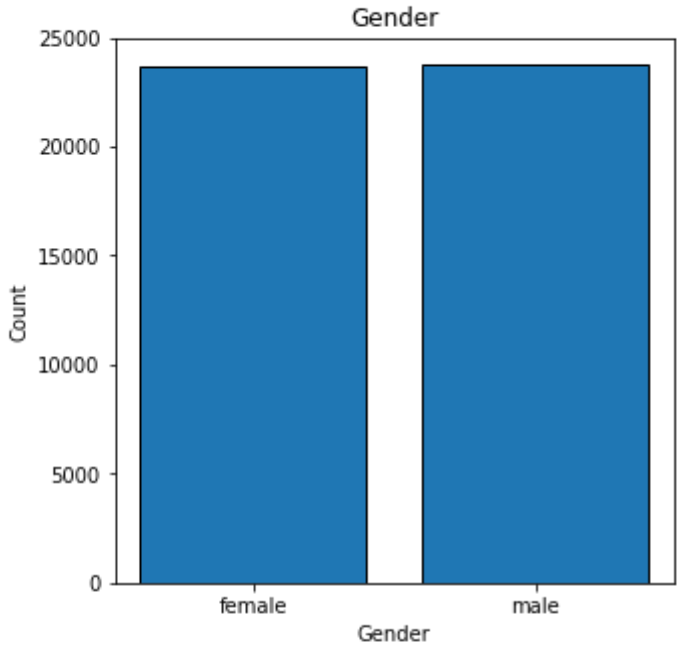
In [13]:
```
# from scipy.stats import boxcox

# df.purch_amt = boxcox(df.purch_amt)[0]
# df.age = boxcox(df.age)[0]
# df.days_since_last_purch = boxcox(df.days_since_last_purch)[0]
```

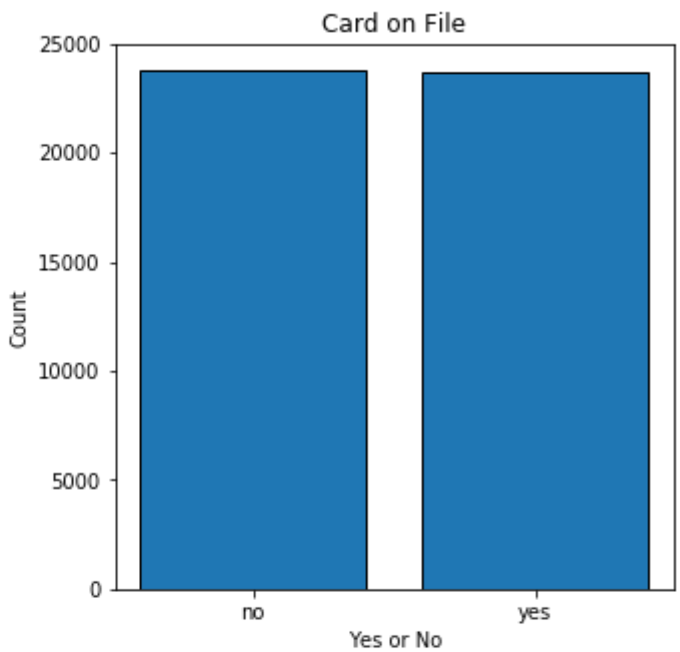
In [14]:
```
df.head()
```

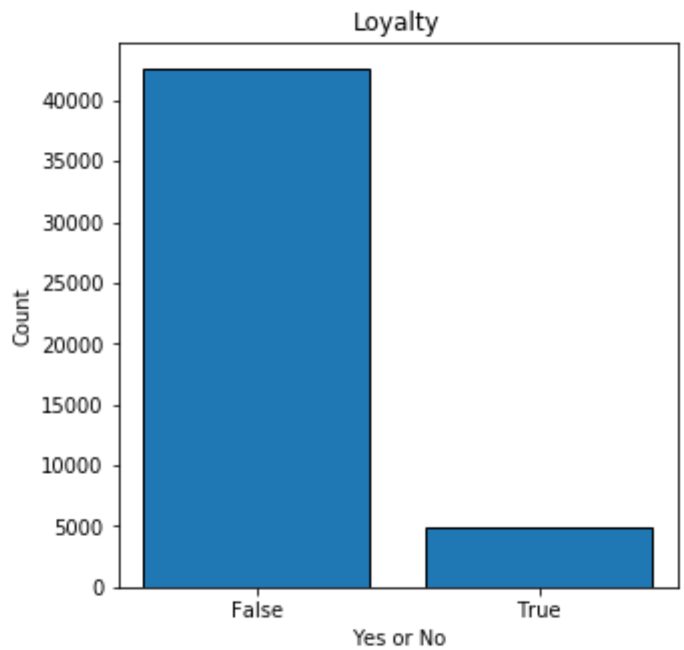
Out[14]:

	purch_amt	gender	card_on_file	age	days_since_last_purch	loyalty
1	65.16	male	yes	23.0	61.0	False
2	40.60	female	no	36.0	49.0	False
3	38.01	male	yes	47.0	57.0	False
6	43.96	male	yes	36.0	64.0	False
9	93.63	female	no	40.0	47.0	True

In [15]:
```
# Visualize categorical variables, class balances

x = ['female', 'male']
y = list(df.groupby(['gender']).count().loyalty)

plt.figure(figsize=(5,5))
plt.bar(x,y, edgecolor = 'k')
plt.title('Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

x = ['no', 'yes']
y = list(df.groupby(['card_on_file']).count().loyalty)

plt.figure(figsize=(5,5))
plt.bar(x,y, edgecolor = 'k')
plt.title('Card on File')
plt.xlabel('Yes or No')
plt.ylabel('Count')
plt.show()

x = ['False', 'True']
y = list(df.groupby(['loyalty']).count().gender)

plt.figure(figsize=(5,5))
plt.bar(x,y, edgecolor = 'k')
plt.title('Loyalty')
plt.xlabel('Yes or No')
plt.ylabel('Count')
plt.show()
```


Gender
25000
20000
15000
10000
5000
0
Count
female
male
Gender


Card on File
25000
20000
15000
10000
5000
0
Count
no
yes
Yes or No



Correct Class Imbalance

In [16]:
```
# Correct Class Imbalance by Under Sampling "No Loyalty"
no_loyalty = df[df['loyalty'] == False]
no_loyalty = no_loyalty.sample(frac=.115)

loyalty = df[df['loyalty'] == True]

undersample = pd.concat([no_loyalty, loyalty])
undersample = undersample.reset_index()
undersample = undersample.drop(['index'], axis=1)
```

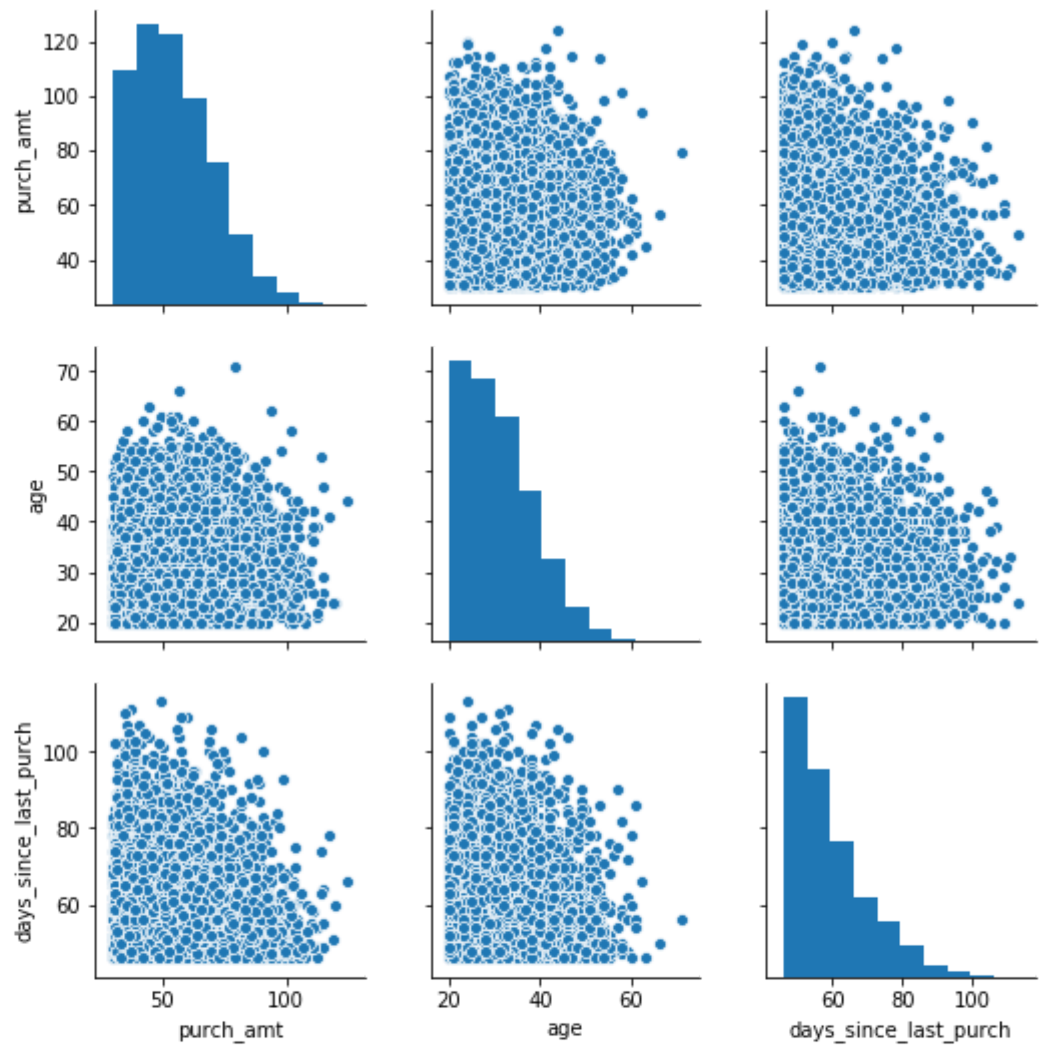
In [17]:
```
undersample.shape
```

Out[17]:
```
(9741, 6)
```

In [18]:
```
# Visualize Class Balabnce of Loyalty
x = ['False', 'True']
y = list(undersample.groupby(['loyalty']).count().gender)

plt.figure(figsize=(6,6))
plt.bar(x,y, edgecolor = 'k')
plt.title('Loyalty')
plt.xlabel('Yes or No')
plt.ylabel('Count')
plt.show()
```

In [19]:
```
# Visualize Correlation and Distributions of undersampled dataset
sns.pairplot(undersample[['purch_amt', 'age', 'days_since_last_purch']])
```

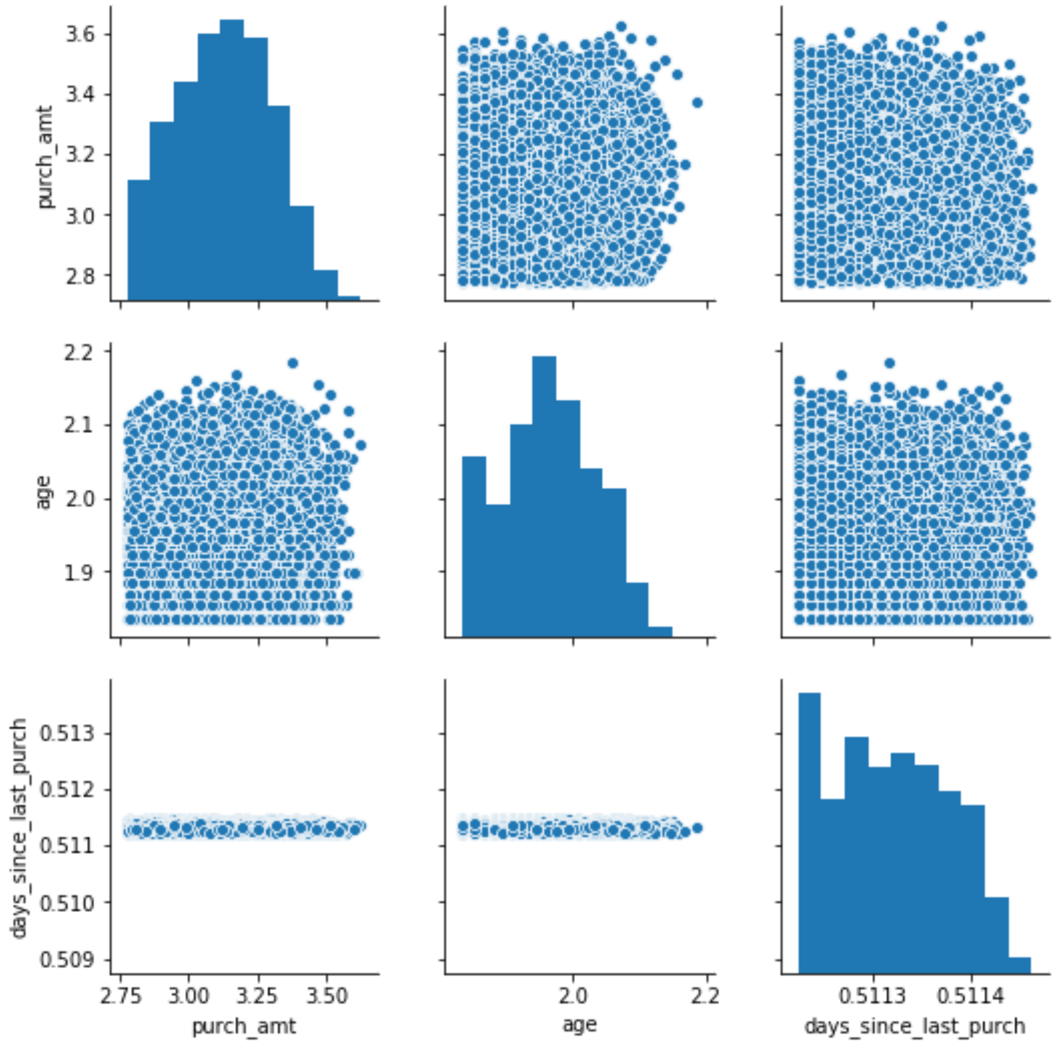
Out[19]:
```
<seaborn.axisgrid.PairGrid at 0x1223a2320>
```



In [20]:
```
# Correct Right Skew distribution with Box Box Distribution
undersample.purch_amt = boxcox(undersample.purch_amt)[0]
undersample.age = boxcox(undersample.age)[0]
undersample.days_since_last_purch = boxcox(undersample.days_since_last_p
urch)[0]
```

In [21]:
```
# Visualize Distributions after BoxCox Transformation
sns.pairplot(undersample[['purch_amt', 'age', 'days_since_last_purch']])
```

Out[21]: <seaborn.axisgrid.PairGrid at 0x1223a23c8>



In [22]:
```
# Turn categorical Variables in to Numeric
undersample = pd.get_dummies(undersample, columns = ['gender', 'card_on_
file'])
undersample['loyalty'] = np.where(undersample['loyalty']==True, 1, 0)
```

In [23]:
```
undersample.head()
```

Out[23]:

	purch_amt	age	days_since_last_purch	loyalty	gender_female	gender_male	card_on_file_
0	3.115137	1.868913	0.511267	0	1	0	
1	3.308758	1.965896	0.511346	0	0	1	
2	3.217503	2.018075	0.511335	0	0	1	
3	3.315823	2.059608	0.511302	0	1	0	
4	3.181516	1.934152	0.511309	0	0	1	

Baseline Models

In [24]:
```
# Split into test and training set
X = undersample.drop(['loyalty'], axis=1)
Y = undersample['loyalty']
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.2,
random_state=0)

# List of models
models = [RandomForestClassifier(n_estimators=200, max_depth=3, random_s
tate=42), LinearSVC(),
          GaussianNB(), LogisticRegression(random_state=42), GradientBoo
stingClassifier(n_estimators=200, random_state=42)]

model_names = []
accuracies = []

# Iterate through models to compare accuracies
for model in models:
    model_name = model.__class__.__name__
    model_names.append(model_name)
    model.fit(x_train, y_train)
    predictions = model.score(x_train, y_train)
    accuracies.append(predictions)


# Assign accuracies for each model to view
accuracy_df = pd.DataFrame()
accuracy_df['Model'] = model_names
accuracy_df['Accuracy'] = accuracies
```

```
/usr/local/lib/python3.6/site-packages/sklearn/linear_model/logistic.p
y:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.2
2. Specify a solver to silence this warning.
  FutureWarning)
```

In [25]:
```
accuracy_df
```

Out[25]:

	Model	Accuracy
0	RandomForestClassifier	0.727156
1	LinearSVC	0.614861
2	GaussianNB	0.739861
3	LogisticRegression	0.612038
4	GradientBoostingClassifier	0.764759

In [26]:
```
# GradientBoostingClassifier cross validation

gbc = GradientBoostingClassifier(n_estimators=200, random_state=42)
gbc.fit(x_train, y_train)
y_pred = gbc.predict(x_test)
cross_val_score(gbc, X, Y, cv=5)
```

Out[26]:
```
array([0.73114418, 0.73921971, 0.74075975, 0.7412731 , 0.71868583])
```

In [27]:
```
# View accuracy scores on classifying each author (precission, recall, f
1-score and support)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.75      0.69      0.72       957
           1       0.72      0.78      0.75       992

   micro avg       0.74      0.74      0.74      1949
   macro avg       0.74      0.73      0.73      1949
weighted avg       0.74      0.74      0.73      1949
```

In [28]:
```
# Pass linear regression model to the RFE constructor

selector = RFE(gbc)
selector = selector.fit(X, Y)

# Sort ranked features
pd.set_option("display.max_rows", 999)
rankings = pd.DataFrame({'Features': X.columns, 'Ranking' : selector.ran
king_})
rankings = rankings.sort_values('Ranking')
rankings = rankings.reset_index()
rankings = rankings.drop(columns=['index'])
rankings = rankings.set_index('Features')
rankings = rankings.T
rankings
```

Out[28]:

Features	purch_amt	age	days_since_last_purch	card_on_file_no	gender_male	card_on_file_yes
Ranking	1	1	1	2	3	4

In [29]:
```
# RandomForestClassifier cross validation

rfc = RandomForestClassifier(n_estimators=300, max_depth=2, random_state
=42)
rfc.fit(x_train, y_train)
y_pred = rfc.predict(x_test)
cross_val_score(rfc, X, Y, cv=5)
```

Out[29]:
```
array([0.72652642, 0.73100616, 0.71201232, 0.72022587, 0.70174538])
```

In [30]:
```
# View accuracy scores on classifying each author (precission, recall, f
1-score and support)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.78      0.57      0.66       957
           1       0.67      0.85      0.75       992

   micro avg       0.71      0.71      0.71      1949
   macro avg       0.73      0.71      0.70      1949
weighted avg       0.73      0.71      0.71      1949
```

In [31]:
```
# Pass linear regression model to the RFE constructor

selector = RFE(rfc)
selector = selector.fit(X, Y)

# Sort ranked features
pd.set_option("display.max_rows", 999)
rankings = pd.DataFrame({'Features': X.columns, 'Ranking' : selector.ran
king_})
rankings = rankings.sort_values('Ranking')
rankings = rankings.reset_index()
rankings = rankings.drop(columns=['index'])
rankings = rankings.set_index('Features')
rankings = rankings.T
rankings
```

Out[31]:

Features	purch_amt	age	days_since_last_purch	card_on_file_no	gender_male	gender_female	c
Ranking	1	1	1	2	3	4	

Feature Engineering

In [32]:
```
# Feature Engineering
undersample['purch_amt'] = undersample['purch_amt'].apply(lambda x: x**2
)
undersample = undersample.drop(['card_on_file_yes', 'gender_male'], axis
=1)
```

Second Round Training on New Features and Predictions

In [33]:
```
# GradientBoostingClassifier cross validation

gbc = GradientBoostingClassifier(n_estimators=200, random_state=42)
gbc.fit(x_train, y_train)
y_pred = gbc.predict(x_test)
cross_val_score(gbc, X, Y, cv=5)
```

Out[33]:
```
array([0.73114418, 0.73921971, 0.74075975, 0.7412731 , 0.71868583])
```

In [34]:
```
# View accuracy scores on classifying each author (precission, recall, f
1-score and support)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.75      0.69      0.72       957
           1       0.72      0.78      0.75       992

   micro avg       0.74      0.74      0.74      1949
   macro avg       0.74      0.73      0.73      1949
weighted avg       0.74      0.74      0.73      1949
```

In [35]:
```
# RandomForestClassifier cross validation

rfc = RandomForestClassifier(n_estimators=300, max_depth=2, random_state
=42)
rfc.fit(x_train, y_train)
y_pred = rfc.predict(x_test)
cross_val_score(rfc, X, Y, cv=5)
```

Out[35]:
```
array([0.72652642, 0.73100616, 0.71201232, 0.72022587, 0.70174538])
```

In [36]:
```
# View accuracy scores on classifying each author (precission, recall, f
1-score and support)
print(metrics.classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.78      0.57      0.66       957
           1       0.67      0.85      0.75       992

   micro avg       0.71      0.71      0.71      1949
   macro avg       0.73      0.71      0.70      1949
weighted avg       0.73      0.71      0.71      1949
```

In [ ]: