Final Project

Statistical Prediction Modeling *DATA 2204*

Jumana Obeid

100832624

DURHAM COLLEGE
SUCCESS MATTERS

Rational Statement (summary of the problem or problems to be addressed by the PPT) Slide 2-5 (7%)

Mr.Hughes, In this problem, needs to predict if these items are more likely to be spent in one of two “Channels” based on some input, the algorithm that we will use is "Logistic regression”first, this algorithm will return the likelihood(probability) that these input will be spent in one of two (Cafe/Restaurant/Hotel)Chanel or Retail Channel.This type of analysis can help you predict the likelihood of an event happening or a choice being made.This algorithm assumes that the output , follow a binomial distribution, and that there is a linear relationship between the independent variables and the link function(logit).Also , the independent variables(input/features) have mutually exclusive and exhaustive categories.

Another algorithm is Random Forest, consists of a large number of individual decision trees that operate as an ensemble. spits out a class prediction and the class with the most votes becomes our model's prediction either (Cafe/Restaurant/Hotel)Chanel or Retail Channel.

The independent variables are (the features):

Independent Variables:
Fresh: annual spending ($000) on fresh products
Milk: annual spending ($000) on milk products
Grocery: annual spending ($000) on grocery products
Frozen: annual spending ($000) on frozen products Detergents_
Paper: annual spending ($000) on detergents and paper products
Delicatessen: annual spending ($000) on and delicatessen products
Region: Lisbon (1), Oporto (2) or Other (3)
Dependent Variables:Channel: Hotel/Restaurant/Café(1) or Retail channel (2)

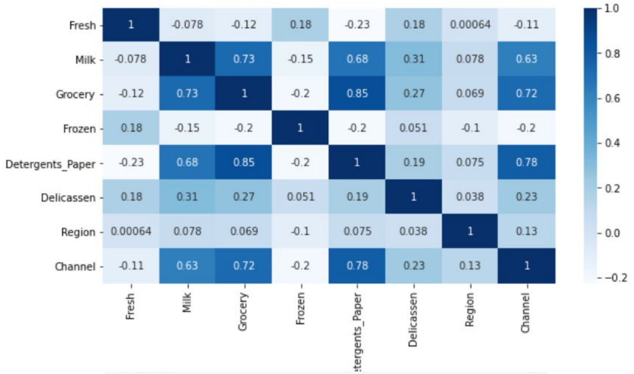
Explain feature engineering techniques (in your own words) that were used to clean and prepare the dataset (i.e. Tukey, SMOTE, and SelectFromModel).

Tukey removes outliers in a data set. It first calculates the first quartile, the second quartile and the interquartile. Then it calculates upper boundary and lower boundaries based on the interquartile. Any value that is less than the lower boundary or any value that is higher than upper boundary is removed from the dataset rows.

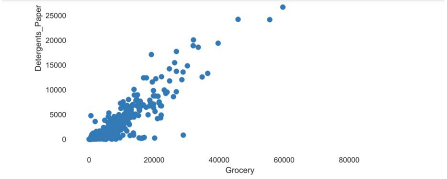
Smote balances the dataset, it adds points to the class with less values, so the classifiers are balanced , having almost the same number of data points,for better predictions.

SelectFromModel, selects the critical features considered vital input for the prediction.

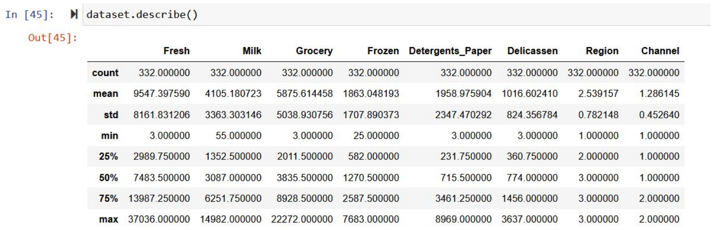
- **Using high-level exploratory data analysis (EDA), identify and explain three (3) key insights from the ChannelDataset.csv dataset. For example: Pandas Profiling, .Describe(), Correlation Matrix, etc**.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region	Channel
Fresh	1	-0.078	-0.12	0.18	-0.23	0.18	0.00064	-0.11
Milk	-0.078	1	0.73	-0.15	0.68	0.31	0.078	0.63
Grocery	-0.12	0.73	1	-0.2	0.85	0.27	0.069	0.72
Frozen	0.18	-0.15	-0.2	1	-0.2	0.051	-0.1	-0.2
Detergents_Paper	-0.23	0.68	0.85	-0.2	1	0.19	0.075	0.78
Delicassen	0.18	0.31	0.27	0.051	0.19	1	0.038	0.23
Region	0.00064	0.078	0.069	-0.1	0.075	0.038	1	0.13
Channel	-0.11	0.63	0.72	-0.2	0.78	0.23	0.13	1



3- From Pandas Profiling and heat map, we can see a high positive correlation between Grocery and Detergents_Paper.(80%).



In [45]:
```
dataset.describe()
```

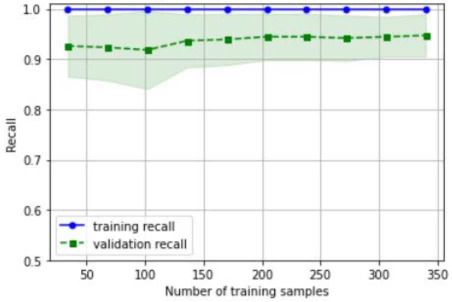
Out[45]:

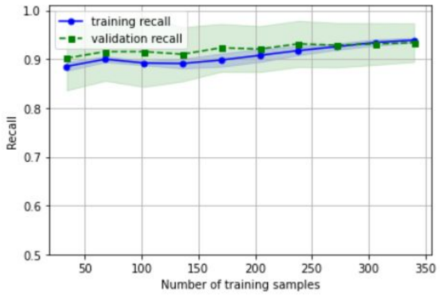
	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Region	Channel
count	332.000000	332.000000	332.000000	332.000000	332.000000	332.000000	332.000000	332.000000
mean	9547.397590	4105.180723	5875.614458	1863.048193	1958.975904	1016.602410	2.539157	1.286145
std	8161.831206	3363.303146	5038.930756	1707.890373	2347.470292	824.356784	0.782148	0.452640
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000	1.000000	1.000000
25%	2989.750000	1352.500000	2011.500000	582.000000	231.750000	360.750000	2.000000	1.000000
50%	7483.500000	3087.000000	3835.500000	1270.500000	715.500000	774.000000	3.000000	1.000000
75%	13987.250000	6251.750000	8928.500000	2587.500000	3461.250000	1456.000000	3.000000	2.000000
max	37036.000000	14982.000000	22272.000000	7683.000000	8969.000000	3637.000000	3.000000	2.000000

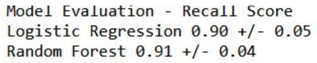
1- Avg spending on milk is 4105$.And we can say that there is a positive correlation between spending on Milk and Channel. (63%).

2-Avg spending on Fresh items is 9547$, almost double the spending on milk, and it has a low correlation with the Channel. In other words, there is spending on Fresh items in all Channels.(20%)

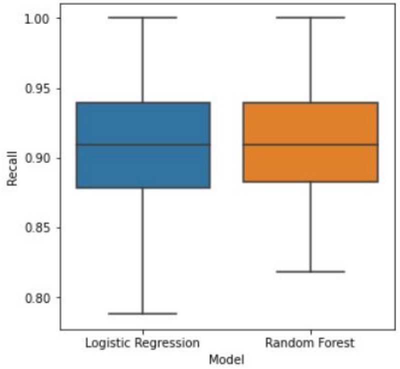
Present and explain three (3) insights from the Learning Curve of both algorithms. Please use accuracy for your scoring (i.e. scoring='accuracy')(1)


Random Forest


Logistic Regression Learning Curve



```
Model Evaluation - Recall Score
Logistic Regression 0.90 +/- 0.05
Random Forest 0.91 +/- 0.04
```


Boxplot View

1-As we can see from the learning curve, there is high variance.(overfitting)

2-Recall tells us that 90% were correct in Logistic regression out of all truths.

3-Recall for a random forest is 91%.

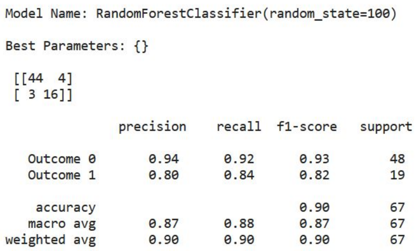
Present and explain three (3) insights from the Learning Curve of both algorithms. Please use accuracy for your scoring (i.e. scoring='accuracy')(2)

Random Forest

F1 score is the harmonic mean between Recall and Precision and is 90%

Recall is out of all truths we got 90%

Precision is out of all predictions we got 90%



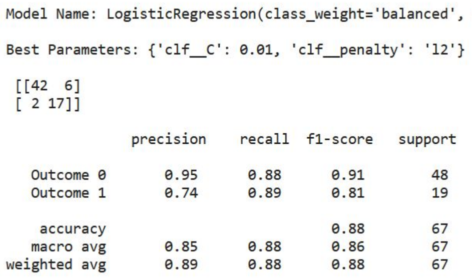
```
Model Name: RandomForestClassifier(random_state=100)

Best Parameters: {}

 [[44  4]
 [ 3 16]]
              precision    recall  f1-score   support

   Outcome 0       0.94      0.92      0.93        48
   Outcome 1       0.80      0.84      0.82        19

    accuracy                           0.90        67
   macro avg       0.87      0.88      0.87        67
weighted avg       0.90      0.90      0.90        67
```

Logistic Regression

F1 score is the harmonic mean between Recall and Precision and is 88%

Recall is out of all truths we got 88%

Precision is out of all predictions we got 89%



```
Model Name: LogisticRegression(class_weight='balanced',

Best Parameters: {'clf__C': 0.01, 'clf__penalty': 'l2'}

 [[42  6]
 [ 2 17]]
              precision    recall  f1-score   support

   Outcome 0       0.95      0.88      0.91        48
   Outcome 1       0.74      0.89      0.81        19

    accuracy                           0.88        67
   macro avg       0.85      0.88      0.86        67
weighted avg       0.89      0.88      0.88        67
```

• Explain the models (in your own words) that were used to help predict channel classification. Remember that you must only use: • Logistic Regression • Random Forest

The first algorithm is "**Logistic regression**” this algorithm will return the probability of a having either a (Cafe/Restaurant/Hotel)Channel or Retail Channel.
This type of analysis can help you predict the likelihood of an event happening or a choice being made.This algorithm assumes that the output , follow a binomial distribution, and that there is a linear relationship between the independent variables and the link function(logit).

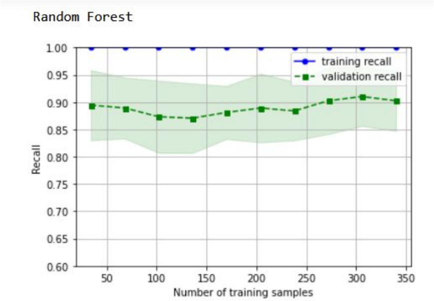
The other algorithm is “**Random Forest”**, Random forest is composed of Decision Trees,partitioning carried out by a Classification Tree, the output is one of two classes.classification tree, you predict that each observation belongs to the most occurring class and we are interested in the branches as well.
In a **Random Forest**:

1. Many decision trees are trained,but it creates several subsets of data in each tree.
2. Each node only considers a subset of features in every tree.

Present and Explain three (3) key insights from the classification report metrics for the of each of the two optimized models (i.e. Logistical Regression and Random Forest). Note: a total of six (6) are required.(1)After Feature Selection

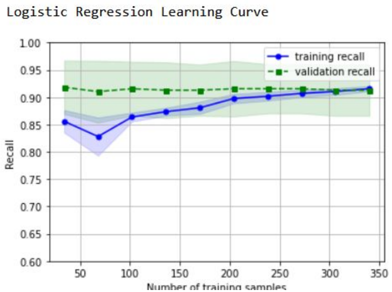
Random Forest

From the learning curve we can say that the logistic regression has high variance(overfitting).



Logistic Regression

From the learning curve we can say that the logistic regression has high variance(overfitting).



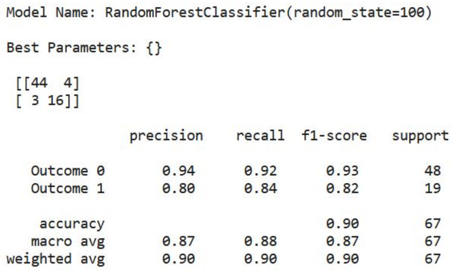
Present and Explain three (3) key insights from the classification report metrics for the of each of the two optimized models (i.e. Logistical Regression and Random Forest). Note: a total of six (6) are required.(2)After Feature Selection

Random Forest

F1 score is the harmonic mean between Recall and Precision and is 90%

Recall is out of all truths we got 90%

Precision is out of all predictions we got 90%



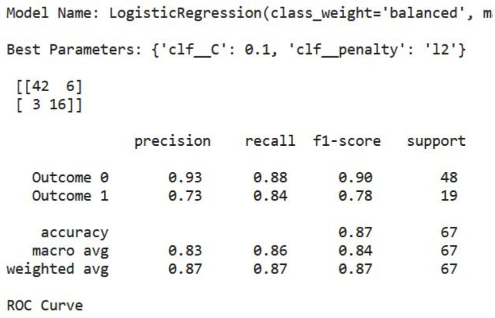
```
Model Name: RandomForestClassifier(random_state=100)

Best Parameters: {}

 [[44  4]
 [ 3 16]]

              precision    recall  f1-score   support

   Outcome 0       0.94      0.92      0.93        48
   Outcome 1       0.80      0.84      0.82        19

    accuracy                           0.90        67
   macro avg       0.87      0.88      0.87        67
weighted avg       0.90      0.90      0.90        67
```

Logistic Regression

F1 score is the harmonic mean between Recall and Precision and is 87%

Recall is out of all truths we got 87%

Precision is out of all predictions we got 87%



```
Model Name: LogisticRegression(class_weight='balanced', m

Best Parameters: {'clf__C': 0.1, 'clf__penalty': 'l2'}

 [[42  6]
 [ 3 16]]

              precision    recall  f1-score   support

   Outcome 0       0.93      0.88      0.90        48
   Outcome 1       0.73      0.84      0.78        19

    accuracy                           0.87        67
   macro avg       0.83      0.86      0.84        67
weighted avg       0.87      0.87      0.87        67

ROC Curve
```

Present and Explain two (2) key insights from ROC/AUC Curve (Optimized Model)

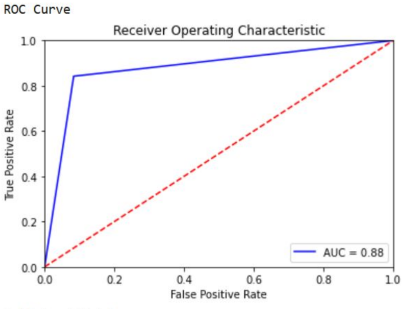
AUC is 88%, (Area under the curve).is very good the higher the better.

ROC plot , shows the true positive rate to false positive rate.The closer to right angle the better,.

A true positive is **an outcome where the model correctly predicts the positive class**.

A false positive outcome where the model incorrectly predicts the positive class

Random Forest



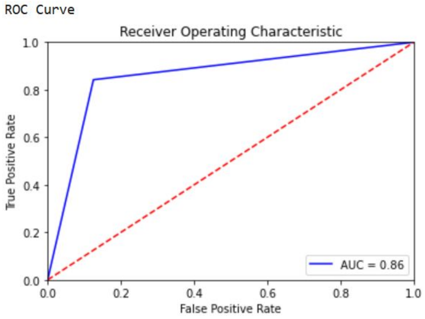
AUC is 86%, (Area under the curve).is very good the higher the better.

ROC plot , shows the true positive rate to false positive rate.The closer to right angle the better,

A true positive is **an outcome where the model correctly predicts the positive class**.

A false positive outcome where the model incorrectly predicts the positive class

Logistic Regression

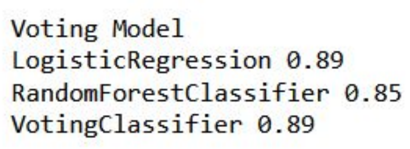


Compare the Optimized Logistic Regression model to the Optimized Random Forest model identifying two (2) key insights

1-Random Forest performed better than Logistic regression,
F1 score for Random Forest was, 90%.
While F1 score for Logistic Regression is 88%.
After selecting important features , F1 score for Random Forest remained 90%.
But F1 score for Logistic Regression was 87%.

2- Studying ROC, AUC , Precision and Recall.Random Forest performed better than Logistic Regression.

• Explain your rationale in developing the Ensemble Voting Model (i.e. how did you decide on the algorithm and technique)(1)



```
Voting Model
LogisticRegression 0.89
RandomForestClassifier 0.85
VotingClassifier 0.89
```

Voting is a simple algorithm. It works by first creating two or more standalone models from your training dataset.

A Voting Classifier can then be used to wrap models and average the predictions of the sub-models when asked to make predictions for new data.

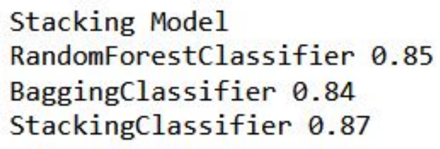
In other words, Ensemble is the art of combining diverse set of learners (individual models) together to improvise

on the stability and predictive power of the model.The voting classifier score is 89%.

• Explain your rationale in developing the Ensemble Voting Model (i.e. how did you decide on the algorithm and technique)(2)

I used Bagging

1. repeatedly randomly resampling the training data
2. parallel ensemble: each model is built independently
3. aim to decrease variance, not bias
4. suitable for high variance low bias models (complex models)
5. an example of a tree based method is random forest, which develop fully grown trees (note that RF modifies the grown procedure



```
Stacking Model
RandomForestClassifier 0.85
BaggingClassifier 0.84
StackingClassifier 0.87
```

Stacked

a new model is trained to combine the predictions from two or more models already trained or your dataset.

After blending the models together, combine the predictions for submodels to voting to a weighted sum using linear regression or logistic regression

State your conclusions based on your analysis and provide your recommendation on which model Mr. John Hughes should implement. Don’t forget to justify your answers.

- **Identify and justify two (2) next steps that could be used to help enhance the usability of the model.**

The two algorithms performance is very good, the F1 score for Logistic regression was 87% and the F1 score for Random Forest is 90%, however , there could be an issue of overfitting, **1-** selecting important features, helped a little ,but

2- adding more data and **3-** increasing regularization can be beneficial to reduce this over fitting.