ACS61013 Data Modelling and Machine Intelligence Course work 2 Assignment

Saba Firdaus Ansaria
Registration Number: 210110201

January 3, 2023

Contents

1 Abstract

Data analysis is an essential tool for various business and day to day tasks, which is data-centric learning and decision making. Managing data is crucial to avoid inherent bias and security issues in the underlying model, which can cause significant damage to the model outcome. Hence, data analysis is a compilation of a few sequential structured processes such as data collection, data processing, bias removal, feature engineering, model training, regularisation, filtering etc. The whole process can be simply referred to as a data pipeline. In this report, the use of the data pipeline and its individual stages are discussed and explored through a dataset. Furthermore, the report develops machine learning models to predict customer satisfaction based on airport features.

2 Introduction

A machine learning pipeline is a structured assemblage of processes consisting of a set of machine learning algorithms chosen based on their nuances and the data they will process. The complexity of a pipeline depends on the data and the given task. Different data have different biases and risks associated with them. For example, health data consists of people's identities and health conditions that the insurance companies can exploit. The phone recording data consists of ATM numbers and sensitive financial information. Thus, an effective machine learning pipeline should remove as much bias and sensitive information so that the sensitive information becomes hidden. The features in the data may be biased as well. For example, Twitter data from different regions can have different gender and race bias.

In this report I explored the airport dataset and understand its domain, characteristics and peculiarities (Section 3, 4. Then the data is preprocessed based on the domain understanding, and irrelevantbiased information is filtered (Section 5. After the preprocessing, the correlation among the features is examined using PCA and domain knowledge to understand the data better and separate the outliers (Section 6. The feature engineering leads to the modelling part where machine learning models are trained with the data (Section 7. The machine learning models are analysed & evaluated and necessary measures are taken to limit overfitting and underfitting condition (Section 8, 9). Finally, the models are compared with the other two machine learning models and presented their similarities, differences, advantages and disadvantages (Section 10). Finally, I drew a conclusion based on the experimental evidence I had in the previous sections.

3 Dataset

The dataset is an airport facility dataset that has different airport attributes and amenities as features with their corresponding categorical and continuous feature values. The dataset contains 3502 observations and 37 features.

4 Domain Analysis (Level 1)

Conduct a domain analysis and present your findings as related to the domain of the coursework. Discuss how what you have found from your domain analysis will support and be carried over to other parts of your coursework.

The Orange framework has been used to import the data and visualise the distribution and correlations. The *File* widget and the *Distributions* widget have been used. The data set describes the quantities and qualitative attributes of airport facilities. A typical traveller has concerns with the attributes of an airport that influence their opinion about the airport, such as

- How is the transfer service from one terminal to another terminal?
- How informative are the airport systems, such as connection timings and flight announcement monitors?
- How are the hygiene amenities (toilets, bathrooms, water supply, pool, cleaning services etc)?
- How many restaurants and eating places are there and the type of cuisine?
- Baggage transfer average time and delays
- How are the airport check-in system and security checking process?
- Is there a garage or parking space, and what is the condition?
- What is the total area and the overall condition of the airport?

The data set contains 37 variable attributes such as Quarter (of the year), Date recorded Date & Time Departure time, Ground transportation to/from airport, Parking facilities, Parking facilities (value for money), Availability of baggage carts, Efficiency of check-in staff , Check-in wait time, Courtesy of check-in staff, Wait time at passport inspection, Courtesy of inspection staff, Courtesy of security staff, Thoroughness of security inspection, Wait time of security inspection, Feeling of safety and security, Ease of finding your way through the airport, Flight information screens, Walking distance inside terminal, Ease of making connections, Courtesy of airport staff, Restaurants, Restaurants (value for money), Availability of banks/ATM/money changing, Shopping facilities, Shopping facilities (value for money), Internet access, Business/executive lounges, Availability of washrooms, Cleanliness of washrooms, Comfort of waiting/gate areas, Cleanliness of airport terminal, Ambience of airport, Arrivals passport and visa inspection, Speed of baggage delivery, Customs inspection, Overall satisfaction. The target variable is 'Overall satisfaction'. Most of the values are continuous numerical values. Clearly, these features are highly relevant to a traveller's overall perception about an airport. An average traveller spends time between his/her connection of flight. The toilets, bathrooms, lounges, and restaurants became very important. Another significant aspect of a travellers concern is the time he/she has to go through the security checking and bagging inspection. The less time it takes, the more favourable the traveller's experience becomes. At the final destination, the time to get the pieces of baggage from the bagging belts is a big concern, and this delay depends on the efficiency of the airport to transfer baggage from the aeroplanes.

The overall ambience of the airport is a significant contributing factor to the customer experience. Be that the availability of additional luxury like having shopping place inside the airport, ATMs and multi-currency exchange availability. The features explain these factors in the dataset. Also, there are some additional critical qualitative features such as the staff behaviour, transport inside the terminals, security and safety. One clear thing

is that the features associated with the customers' well being and customers' time saving will get the highest priority for judging the airports overall quality. However, the time of the year or the time of the flight arrival should not affect the traveler's decision. Because these are not logically related to the airport, neither the airport can control these aspects fully.

Based on this domain knowledge, it would be easier to understand the inter-dependency and correlation of the variables in this data set. The understanding of the domain is helpful to construct an efficient pre-processing process in the machine learning pipeline, which is discussed in the next section (Section 5).

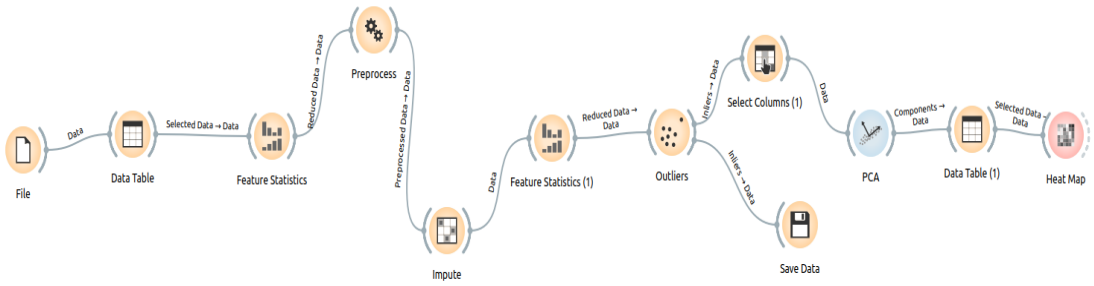


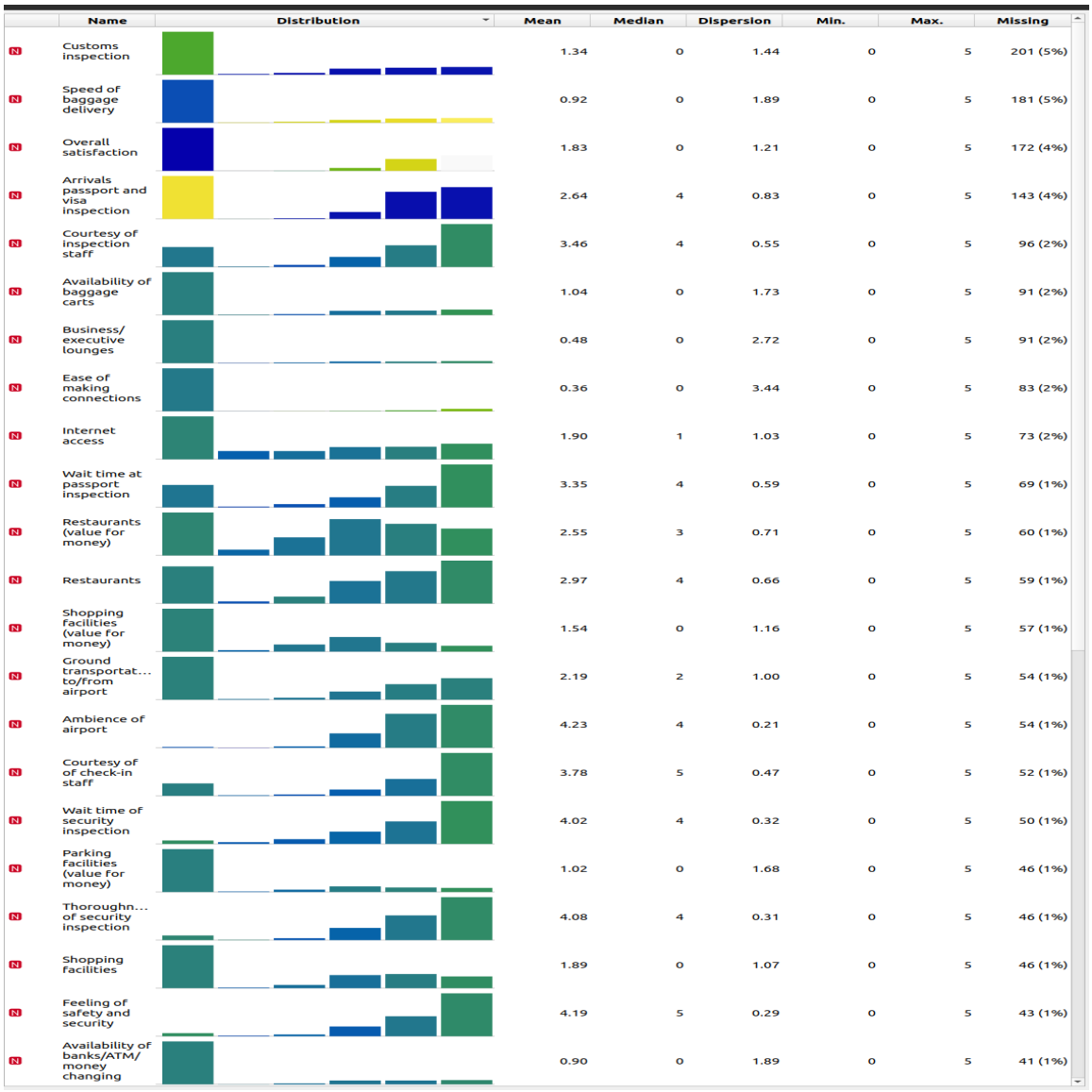
Figure 1: Preprocess Pipeline Orange

5 Data Pre-processing & Feature Engineering (Level 2)

Achieve level 1 as well as conduct data cleaning and pre-processing. Discuss how you used your understanding of the domain from level 1 to support this task.

The data pre-processing is also done using *Orange* framework. The dataset is imported with the file import widget, and the 'Feature Statistics' widget is used to analyse the number of missing features. It has been found that except from the *Availability of washrooms, Cleanliness of washrooms, Comfort of waiting/gate areas, Cleanliness of airport terminal, Ambience of airport, Arrivals passport and visa inspection, Speed of baggage delivery, Customs inspection* features, the rest of the 28 features are missing data. The amount of missing data ranges from 5% to 1%. The target variable *Overall Satisfection* also misses 4% data. This is a tricky situation where most features have some missing data, and the values on these features are not normally distributed. The instances with the missing target variable *Overall Satisfection* are discarded because I do not want to impose unnecessary randomness in the target value by guessing or averaging values in the feature variable. In the case of the rest of the feature variables, the missing values are filled using the average of the overall values of that particular feature variable. Because by removing all those instances with missing values, one-third of the total dataset was about to being discarded. All the feature variables are normalised to standard deviation one and zero mean.

Furthermore, the domain analysis shows that the time of the year or the flight arrival/departure times logically should not influence customers' perception of the airport



Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
Customs inspection		1.34	0	1.44	0	5	201 (5%)
Speed of baggage delivery		0.92	0	1.89	0	5	181 (5%)
Overall satisfaction		1.83	0	1.21	0	5	172 (4%)
Arrivals passport and visa inspection		2.64	4	0.83	0	5	143 (4%)
Courtesy of inspection staff		3.46	4	0.55	0	5	96 (2%)
Availability of baggage carts		1.04	0	1.73	0	5	91 (2%)
Business/ executive lounges		0.48	0	2.72	0	5	91 (2%)
Ease of making connections		0.36	0	3.44	0	5	83 (2%)
Internet access		1.90	1	1.03	0	5	73 (2%)
Wait time at passport inspection		3.35	4	0.59	0	5	69 (1%)
Restaurants (value for money)		2.55	3	0.71	0	5	60 (1%)
Restaurants		2.97	4	0.66	0	5	59 (1%)
Shopping facilities (value for money)		1.54	0	1.16	0	5	57 (1%)
Ground transportat... to/from airport		2.19	2	1.00	0	5	54 (1%)
Ambience of airport		4.23	4	0.21	0	5	54 (1%)
Courtesy of of check-in staff		3.78	5	0.47	0	5	52 (1%)
Wait time of security inspection		4.02	4	0.32	0	5	50 (1%)
Parking facilities (value for money)		1.02	0	1.68	0	5	46 (1%)
Thoroughn... of security inspection		4.08	4	0.31	0	5	46 (1%)
Shopping facilities		1.89	0	1.07	0	5	46 (1%)
Feeling of safety and security		4.19	5	0.29	0	5	43 (1%)
Availability of banks/ATM/ money changing		0.90	0	1.89	0	5	41 (1%)

Figure 2: Histogram of features with missing data

because these parameters are not directly under the airport's control. So the parameters *Quarter (of the year), Date recorded Date & Time Departure time* are removed from the dataset during preprocessing. The 'Preprocess' and 'Impute' widgets have been used to preprocess the data. All the feature variables are turned to continuous variables because the regression process turns them continuous implicitly. This stage is done for better understanding and interpretability of the pipeline. At the end of preprocessing the number of data samples are 3329 and the number of features per data sample is 34.

After this the contradicting and biased samples need to be removed upto a threshold. Euclidean distance has been considered along with 5% contamination threshold while considering 20 neighbour points at the same time. The *Outliers* widget has been used to discard some negative samples. After removing the outliers there are 3187 data samples are left with 34 features variables.

Finally, the airport modelling task can be constructed as both regression task and classification task. If we treat the overall satisfaction as discrete decision variable it can be viewed as classification task otherwise it can be a regression task. In this work both

of them has been investigated.

6 Correlation and components (Level 3)

Achieve the previous levels plus discuss the steps taken in feature engineering (e.g correlation analysis) and preventing bias in the dataset to be used to training the machine learning algorithms. Answer the following questions: Which data features capture the most variability in the dataset and explain why you think they do so? (Hint: Perform PCA first, extract the Principal Components (PCs) that capture the highest variability in the dataset. Then see which features contribute to the PCs). Highlight the PCs together with the features that contribute most to them.

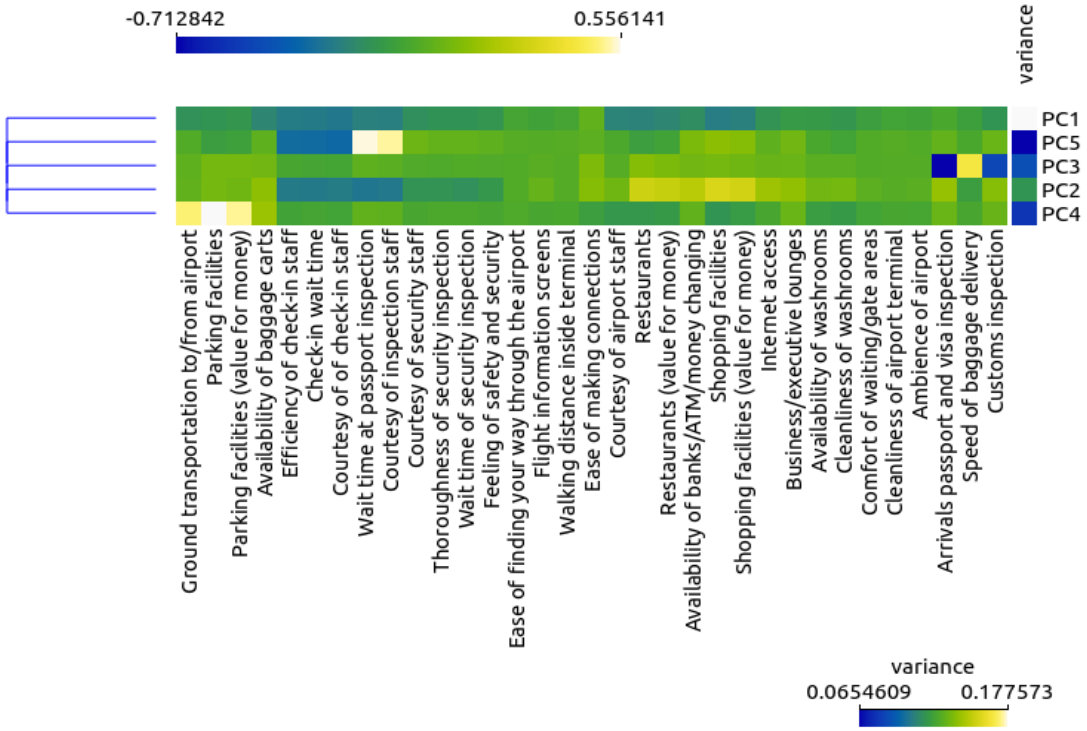


Figure 3: PCA main components and the feature contribution on the components

Principal component analysis is applied on the data. Only 5 components explain 52% of the overall data. Let us call the components as PC1, PC2, PC3, PC4, PC5 and the variance explain ratio is 0.17757268, 0.11554843, 0.08811298, 0.07898028, and, 0.06546079 correspondingly. The rest of the 28 feature variables explain the rest of the variability of the data. The major contributing features on these components are *Cleanliness of airport terminal, Walking distance inside terminal Ease of finding your way through the airport, Arrivals passport and visa inspection, Ease of making connections, Availability of banks/ATM/money changing, Restaurants (value for money), Shopping facilities (value for money), Parking facilities, Speed of baggage delivery, Availability of baggage carts, Ground transportation to/from airport, Parking facilities (value for money), Availability of banks/ATM/money changing, Courtesy of inspection staff, Wait time at passport inspection.* The PCA is done through both the Orange pipeline and python script (both of

Correlation	Feature 1	Feature 2
-0.904	Arrivals passport and visa inspection	Overall satisfaction
0.614	Overall satisfaction	Speed of baggage delivery
-0.507	Customs inspection	Overall satisfaction
0.155	Ease of making connections	Overall satisfaction
0.116	Overall satisfaction	Wait time of security inspection
0.103	Ambience of airport	Overall satisfaction
0.101	Cleanliness of airport terminal	Overall satisfaction
0.094	Overall satisfaction	Wait time at passport inspection
0.09	Feeling of safety and security	Overall satisfaction

Table 1: Correlation Table

which is provided). The variability curve and the components are shown in Figure 3 and 4.

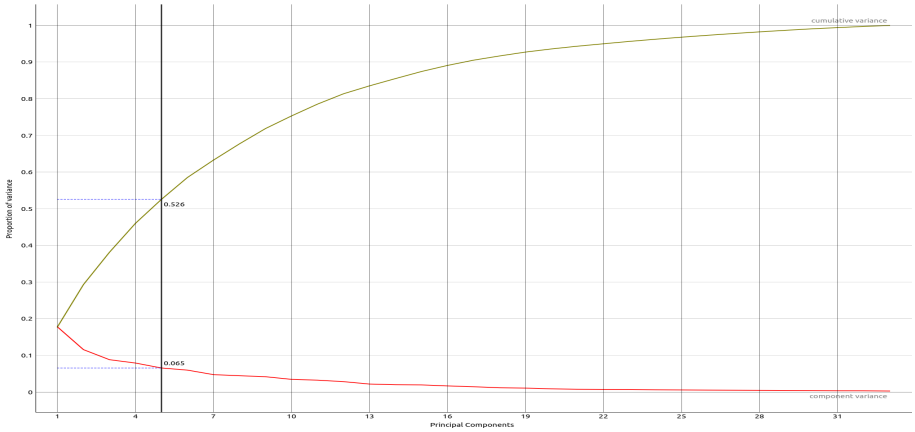


Figure 4: Explained variance by 5 components

The correlation among the features in the data is shown in Table 1. As discussed in Chapter 4, the feature variables are closely related to each other. It is clear that the time each traveller spend inside the airport directly influences his/her perception about the airport. The visa inspection time, customs inspection are negatively correlated. The higher the waiting and processing time, the worse the traveller's perception forms about the airport. The overall cleanliness and ambience of the airport with direct amenities such as restaurants, toilets are highly correlated with the customer satisfaction and overall approval. These finding are logically aligned with the domain analysis from Section 4. As discussed in Section 4, it is clear that a traveller's perception about the airport is directly correlated with the services that the airport provides to reduce traveller's waiting time in different checkpoints (such as security, baggage, passport check etc) and make the transit efficient. The time of the year or the quarter is not logically aligned.

The correlation analysis w.r.t overall satisfaction is done in 'Correlation' widget in Orange pipeline. Pearson correlation is used to calculate the correlation with 'Overall Satisfaction'. The five features such as Arrivals passport and visa inspection, Speed of baggage delivery, Customs inspection, Ease of making connections, Wait time of security inspection hold highest correlation with the target *Overall satisfaction*.

7 Machine Learning Model Pipeline (Level 4)

Achieve all the previous levels as well as explain how:

- **You decided on the choice of the best two machine learning algorithms to apply to the problem.**
- **You used orange (or python/MATLAB) to develop an effective machine learning pipeline from data cleaning up to the point of evaluation.**

The target 'Overall Satisfaction' can be treated as a continuous or categorical discrete variable. Based on the type of the target variable the prediction task can be considered as classification and regression task. In this report, both of the scenarios have been considered. The important question here is to decide which of the two models will be appropriate to train the data based on the sample population and type of data. Two approaches seem more convenient here for the choice of two machine learning model. First one is mixture of experts ensemble approach and the second one is neural network based discriminative model training. Random Forest and Neural Network models have been chosen here for main two machine learning models. Random forests are mixture of decision trees building a superlative decision boundary based on smaller and mixture of expert decision trees. It resample the data over and over and build parallel decision trees. Neural network is a strong model for discriminative training and learning strong mappings between input data and target. The major advantage behind using these models are that they can be used both for classification tasks and regression tasks.

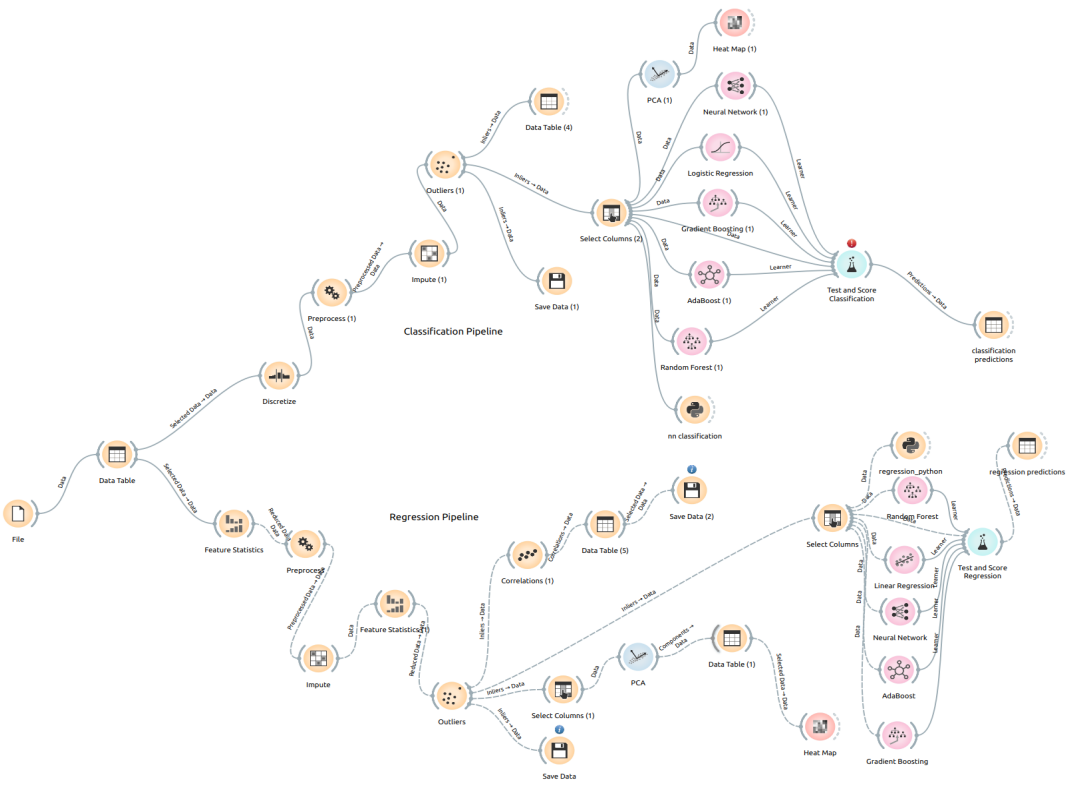


Figure 5: Classification and Regression Pipeline Orange

The Orange framework has been used to build the data pipeline and its widgets to model and train the data. The python widget has also been used to run my python program for training specific regression and classification models.

One thing to notice is that the prepossessing steps are similar in both classification and regression pipelines. The target variable "Overall Satisfaction" is turned into a categorical discrete variable in the classification task. The pipeline is shown in Figure 5. The step by step process of building the pipeline is given below

Step 1: First, add the *File* widget in the current Orange workspace and import the dataset file.

Step 2: View the data on the *Data Table* by adding the widget and connecting the data-flow from the *CSVFile Import* widget.

Step 3: From this point, the data is sent to two parallel pipelines for regression and classification. For the classification pipeline, the "Overall Satisfaction" variable is turned into a categorical discrete variable using the *Discretize* widget. The rest of the steps until model training is similar for both the pipelines.

Step 4: Add the *Feature Statistics* widget and connect the data pipeline from the previous *Data Table* module to view the feature distribution.

Step 5: Open *Feature Statistics* and select all the features except *Departure time, Quarter and Date Recorded* and send them to the next module by clicking on 'Send selection'. By default, it sends automatically. It has been discussed in the domain analysis section why these features are not being considered in this work.

Step 6: Add the *Impute* widget and connect the data pipeline from the previous *Feature Statistics* module. Except for the target variable 'Overall Satisfaction', I have imputed the missing values in all the variables with the 'Average/Most frequent value. The rows with missing target variable values are discarded.

Step 7: Next, the features are normalised with zero mean and standard deviation equal to one. This is done by the *Preprocess* widget.

Step 8: After the preprocessing is done, add the *Outliers* widget to remove some samples with a contamination factor of 5% and consider 20 neighbours for each sample.

Step 9: Add the *Select Columns* widget and select 'Overall Satisfaction' as the target variable. This makes the data ready for the model learning part. Add all the data flow connections shown in Figure 5.

Step 10: Add PCA widget and Heatmap widget to run principal component analysis in the data and then visualise it. This step has been done only in the regression pipeline.

Step 11: Add the model widgets from the menu and Test and Score module as well. Connect the data lines from *Select Columns* to Linear Regression and *Select Columns* to Test and Score.

Step 12: Open Test and Score and select Cross Validation option with 5 folds.

Step 13: Add a Python Script widget and connect the dataflow from *Select Columns*. Now add the python script to run separate regression and classification models to visualise the learning curve and evaluation scores.

Step 14: The models are evaluated using mean square error, precision, recall (Section 8). 5-fold Cross-validation is used both in the Orange Regression model and my python script based regression model. The code is supplied, and the learning curves are shown in Figure 7, 8, 9. The model comparisons are shown in Table 2, 3.

8 Cross validation and Metrics (Level 5)

Achieve all the previous levels plus discuss how you applied cross validation techniques in the machine learning pipeline.

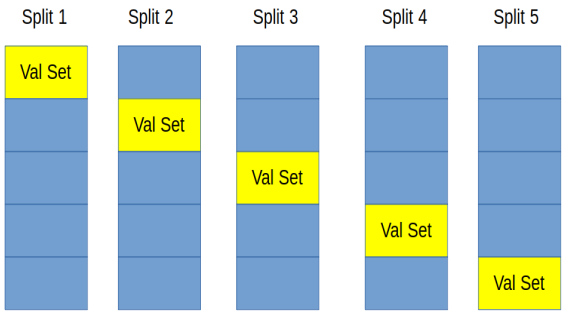


Figure 6: 5-fold cross-validation training/testing regime

In this work, the K-fold cross-validation method has been used to train and test the given dataset. In this scenario k=5 (Figure 6), and the fivefold cross-validation technique chunks the whole sample population into five chunks. Each chunk is left as a test set, and the other four chunks are used as train sets. The train and test sets are divided with 80%/20% ratio. Eventually, in five different iterations, each of the 5 different chunks are used as test sets for once. By this regime, we can use the whole data set while reducing any biases in the data set. As a result, the evaluation gives the whole picture of the data set.

For the regression task validation metric, mean square error (MSE), mean absolute error (MAE) and R^2 value as been used. As for classification task F1 score, precision and recall have been used.

If y is the truth value, $\hat{y}$ is the predicted value, N is the total number of samples and $\bar{y}$ is the mean of y then

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \tag{1}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{2}$$

$$R^2 = 1 - \frac{(y_i - \hat{y})^2}{(y_i - \bar{y})^2} \tag{3}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{4}$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5}$$

$$F1_{Score} = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{6}$$

In the Orange framework Test and Score module has been used and the 5- fold cross validation is applied on that. In the python regression model, the scikit learn cross_validate module has been used.

9 Learning and Tuning (Level 6)

Achieve all the previous levels as well as discuss how effective your pipeline is at preventing overfitting and underfitting through the application of learning curves.

The pipeline deploys a series of precise data preprocessing steps that handle the missing data, bias and data distribution. In section 7, from steps 5-8, the pipeline deals with the missing data and normalisation. The features are normalised with 0 mean and one standard deviation.

The outliers are removed with 20 neighbouring predictions to reduce data bias. Regularisation is used both in the regression task and classification task. The model sizes are chosen according to the data sample poulation size. Generally, bigger models with a huge number of parameters over-fit training when the data sample population is small. For example, the neural network in the classification task has only three hidden layers with 100, 512, 100 parameters, respectively. Adam optimiser is used for optimisation and L1 regularisation is used as well.

I have tried with sample sizes 100, 200, 400, 500, 800, 1000, 1200, 1500, 1800, 2000, 2500 and the learning graph at Figure 9 clearly shows that the higher number helps the training and prevents under-fitting. The PCA reduced dimensions have been used for regression and classification tasks, but they did not significantly improve the results. The random forest and the neural network models both have been tried with different models depths and parameter sizes to overcome overfitting. The random forest is trained with 10 sub-decision trees because our sample population is overall approximately 3000. As the random forest resamples the data repeatedly to build the subtrees, the number of subtrees is 10 to prevent overfitting. The model performance drops by increasing the tree size due to overfitting, and with lesser subtrees, the model performance drops as well due to underfitting.

10 Comparison and Discussion (Level 7)

Achieve all the previous levels and the below:

- You can compare your choice of machine learning algorithm with at least two other algorithms that we have not covered in class.

- Discuss the mathematical peculiarities of the algorithms you have chosen (strengths and weaknesses) and how they impact the results you obtained.
- Apply the appropriate metrics to compare the algorithms you have chosen with the algorithms we have discussed in class.
- Discuss the effects of model complexity of the chosen algorithms on the learning curves generated.

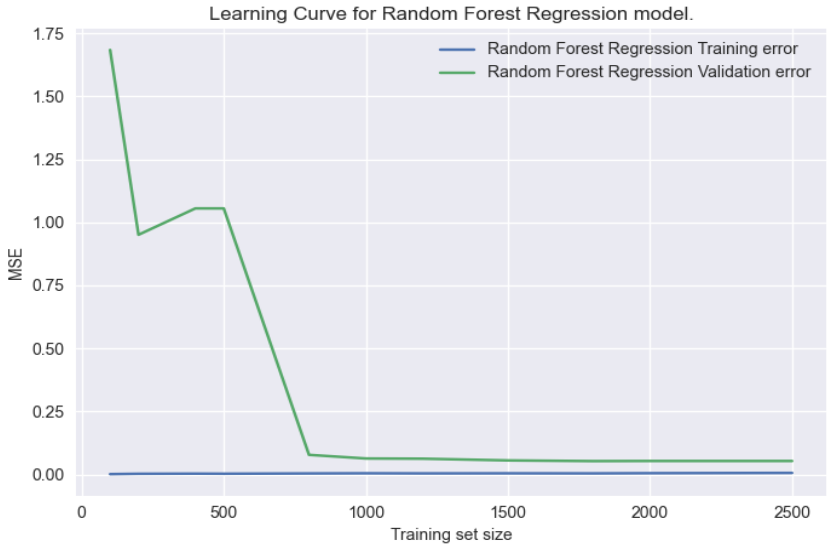


Figure 7: Random Forest Regression Learning Curve Train vs Validation

Figure 8: Adaboost and Gradient boost Regression Learning Curve Train vs Validation

As discussed in Section 7, the main models for the regression and classification task are random forest and neural network. The logic was to choose a model based on mixture

Model	Precision	Recall	F1
Random Forest	0.952	0.959	0.954
Neural Network	0.930	0.937	0.933
Gradient Boost	0.950	0.955	0.952
Ada Boost	0.934	0.930	0.932
Logistic Regression	0.925	0.936	0.929

Table 2: Classification Result Comparisons

Model	MSE	RMSE	MAE	R^2
Random Forest	0.054	0.233	0.089	0.946
Neural Network	0.096	0.311	0.138	0.904
Gradient Boost	0.048	0.219	0.092	0.952
Ada Boost	0.061	0.247	0.069	0.939
Linear Regression	0.127	0.357	0.222	0.872

Table 3: Regression task result comparison

of experts philosophy and a discriminative model. To compare with the chosen models, three more models have been incorporated boosting based models such as ada boosting, gradient boosting and logistic regression, linear regression. Linear regression is used for the regression task and logistic regression has been used for the classification task.

The results have been shown in Table 2 and Table 3. It is quite evident from the tables that the random forest model and gradient boosting model have performed consistently better than the rest of the models. The neural network model is slightly worse than the gradient boost model. The reason here might be due to the smaller number of sample population. The random forest and boosting learning curves are shown in Figure 7 and 8. The boosting algorithms use multiple weak learners and follows the collective decision similarly as the random forests. The random forests also used multiple decision trees with different initialisation and path. This method is generally called mixture of expert approach where the final prediction is an accumulation or weighted average of the collective learners' predictions. The advantage with this approach is its simplicity and interpretability. The linear regression models are the weakest here due to sparsity issue and correlation among the features. Linear regression assumes independence among the feature variables.

Furthermore, the neural network model needs more data than 2500 samples. The balance between sample size, feature variables and model parameter size is very crucial in neural networks. With more number of parameters, the neural network tends to overfit and the F1 score drops to 0.87 and less. Just by adding 2 more hidden layers with layer size 1000 can reproduce this scenario.

The learning curve of linear regression with different objective functions (lasso, ridge) is shown in Figure 9. The Linear regression model is compared with two regression models, which are Adaboost and Gradient Boost regression. Both ada boosting and gradient boosting came from a family of algorithms which is called boosting consisting of a set of weak learners, contrasting with the linear regression where a set of linear coefficients fit the data. In boosting the set of weak learners are trained on the data and then the outcome of those weak learners are averaged in a sequential mixture of experts technique. In adaboost the model weights are re-weighted in iteration and the final weighted output

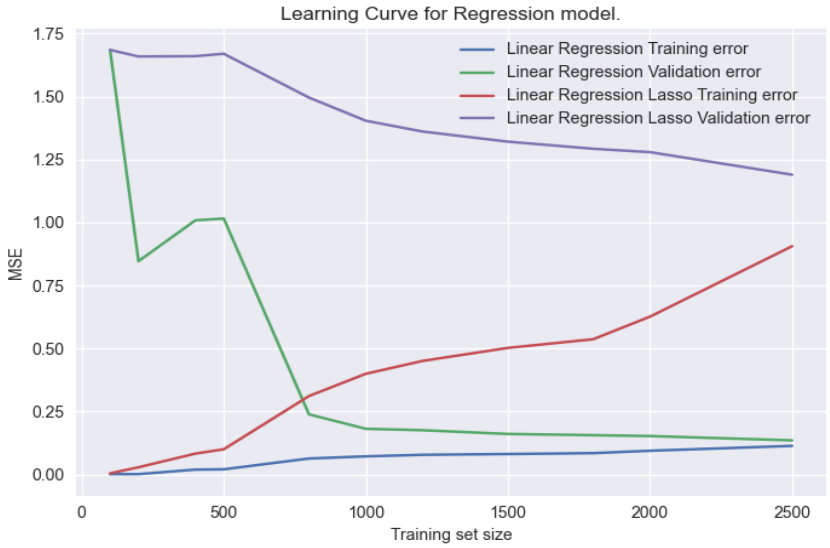


Figure 9: Linear Regression Learning Curve Train vs Validation

of all sub learners decide the final output. Gradient boosting uses gradients to train the sublearners where the loss is back-propagated.

Adaboost is sensitive to outiers where Gradient boost is more precise and robust w.r.t outliers from an optimisation point of view. In Figure 8 it is clearly evident that the model needs more data samples for training. The collection of weak learners and the optimisation strategy makes the boosting algorithms data hungry and they are mostly used with other weak learners. In Figure 8, it is evident that gradient boosting needs more data than ada boosting. It is because of the gradient based optimisation function. However both of these models show high variance problem compare to the linear regression model.

Random forest and boosting models both have decision trees. However, random forest aggregates the sub decision trees at the end, but boosting models aggregate them while learning on the way. In gradient and ada boosting, the sub-learners help other weaker learners to learn and optimise. In forest models, the sub learners learn individually. The weights are interpretable in these approaches compared to the neural network-based models, which are not interpretable straightforwardly.

11 Conclusion

In this report, I have discussed two machine learning data pipelines for the same task. Firstly, the task has been treated as a classification task and then it is treated as a regression task. Random forest, neural network models have been used to train the data. Ada boost, gradient boost, linear regression, logistic regression models have been trained to compare with those two models. The peculiarities and steps of data preprocessing and feature engineering have been discussed with Orange framework and Python programming language. Regression learning has been discussed with learning curves and tables. It is evident that model size is an important factor in controlling model overfitting and underfitting. The model size depends on the training data sample population. The boosting and random forest models have performed significantly better than the other models.

However, the neural network model is not optimised, and it can be further optimised by using different activation functions and hidden layer depth and the number of parameters in each layer. The reason to illustrate airport customer satisfaction modelling as regression and classification models is to explore the possibilities and challenges with discrete and continuous decision boundaries. Both tasks on 'Overall Satisfaction' modelling showed good results with the abovementioned models. However, more complex real-world data is needed to understand the reliability of these models.