Statistical Learning Report

Leonardo Masci

May 26, 2023

1 Exercise 1

1.1 Dataset exploration

The given dataset *NewFlightDelaysTraining.csv* include as indipendent variables: three different arrival airports, John F Kennedy (JFK), Newark (EWR), and LaGuardia (LGA). The three different departure airports are Reagan, Dulles, and Baltimore, along with eight carriers, and a categorical variables for time departure and time scheduled of each flight. The other variables include weather conditions (0 = good/1 = bad) and day of the week (1 for Sunday and Monday and 0 for all other days). The dipendent variable represent if a flight has been delayed by more than 15 min or not (coded as 0 for no delay, and 1 for delay).

Considering that almost the explicative variables are binary, is not usefull to represent the scatter plot of the target variable as a function of each dimension, but is better to convert also the distance and the times of departures in a binary variable, such that the cleaned dataset maintain just 0-1 encoded value. To do that, new columns have been added to the original dataset replacing the *schedtime,deptime,dist*: DelayDep: represent if a flight left on time or not ($delay > 15m$)
LowDist: take into account if the flight exeed 200 miles of distance
MorningFl: it is valued just if a flight has scheduled-time before 14:00 p.m.

To estimate the importance of the variables for the target ones, has been calculated the ratio between how many delayed-flight corrisponds to each valorized value for all the columns, by the relation:

$$importance(i) = \frac{\sum_{i,j}(X_{i,j} \equiv y_i)}{i}$$

and the result is the following:

$varUS$	$varEWR$	$varJFK$	$varLGA$	$varBWI$	$varDCA$
0.653	0.643	0.692	0.460	0.767	0.374

$varCO$	$varDH$	$varDL$	$varMQ$	$varOH$	$varRU$	$varUA$
0.775	0.676	0.664	0.736	0.790	0.699	0.785

$varIAD$	$dayweek$	$DelayDep$	$LowDist$	$MorningFl$	$weather$
0.653	0.737	0.901	0.689	0.462	0.810

The column that represent the *flyghtnumber* has been deleted because is highly correlated with other attributes and is not relevant for the analysis.

Another dataset has been provided, *NewFlightDelaysTest.csv*, and represent the Testing Set for the prediction; it doesn't contain the target variable because has to be estimated for an Aperichallenge, so the original Training Set has been splitted in training_training and training_test.

What is evident from the previous table, is that the most relevant variables is represented by *DelayDep*, that is obvious...if a flight left in delay, has more possibility to lag! The importance is confirmed in the Out of Bag analysis performed in the Bagging Model, how we can see from this picture:

1.2 Statistical Models

Different Statistical Models have been implemented to investigate which gives the best prediction on the Test Set.

According to the expectation, a good model for this kind of data-set could be the **Logistic Regression**; is a statistical model used to analyze the relationship between a categorical dependent variable (binary or multinomial) and one or more independent variables. It estimates the probability of an event occurring by fitting a logistic function to the data. The logistic function, maps the linear combination of predictors to a probability value between 0 and 1. The model calculates odds ratios and estimates the effects of the independent variables on the probability of the outcome. Regressing the data with a binomial distributed logistic curve gives an **AUC**=0.8965.

Another interesting model evaluated is the Naive Bayes approach;

The result is **AUC**=0.847 with a Confusion Matrix:

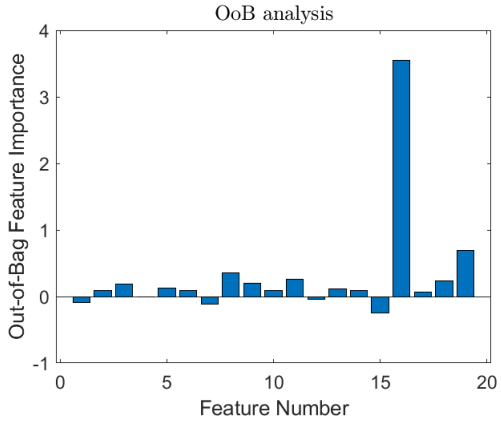
457	12
35	90

A shrinkage method has also been developed to understand better which are the most relevant attributes for predicting delay, such as **LASSO Classification**; the resulting LASSO's coeffincents are the following:

-2.099	2.49	0.9032	-0.477	-0.047	0.000	1.903	0	-0.442	0
-0.694	0.2173	0.222	0	-1.269	0.264	3.882	-2.206	-0.439	5.781

The resulting **AUC**=0.8954

The **Bagging** approach, that involves creating an ensemble of multiple models by training them on different subsets of the original training data, sampled with replacement. In bagging, each model in the ensemble is trained independently, using a different subset of the training data. The predictions from each individual model are then combined through averaging to obtain the final prediction; This model gives as results an **AUC**=0.823 with a relevance of coefficient given by



that evidence the important of the Delay Departure, as predicted in the preprocessing of the dataset.

The **Discriminant Analysis** is a statistical model used to classify observations into predefined groups based on their predictor variables. It aims to find the optimal discriminant function that maximally separates the groups in the feature space. It assumes that the predictors follow a multivariate normal distribution and that the covariance matrices are equal across groups. The model calculates the discriminant scores for each observation and assigns it to the group with the highest score. This method gives as result **AUC**=0.863.

In the end, the last model trained is the **Random Forest**; it is an ensemble learning method that combines multiple decision trees to make predictions. Each tree in the forest is constructed using a random subset of the training data and a random subset of the predictor variables. During prediction, each tree independently provides its output, and the final prediction is determined by averaging across the trees. The application of this Model in this specif case, contradicts in a way the theory, because the variables are all binary, and can be seen as a bad-conditioned case, where splitting the space of the variables leads to overlapping areas. The result however is an **AUC**=0.819

1.3 Results

The result, is in accord with the theoretic prediction, the best model for this kind of dataset is the Logistic Regression, that doesn't bring to overfitting, since is not much complex, and is able to predict with the accuracy of 89% the Delay on the testing set. The confusion matrix reported below can be used to calculate the accuracy of this model by the formula $Precision = TP/(TP + FN)$ that is **PRECISION**=0.83% and an AUC given by the ROC of **AUC**=0.8965

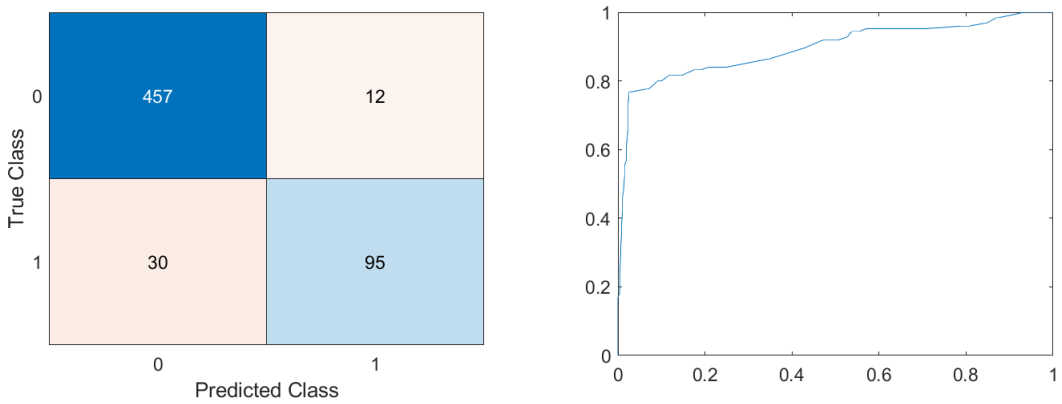


Figure 1: Confusion Matrix and ROC curve for Logistic regression

2 Exercise 2

2.1 Montecarlo Simulations for Statistical model

A normal distributed dataset (with $p = 8$ predictors) has been studied by 200 simulations with different parameters:

- Number of observation $= [100, 500, 1000]$
- Value of ρ between variables in the Covariance Matrix $= [0., 0.5, 0.95]$
- Two different kind of Models has been considered:
  1. **Dense Model** with all predictors relevant $\beta = [1, 1, 1, 1, 1, 1, 1, 1]$
  2. **Sparse Model** with just half predictors relevan $\beta = [1, 1, 1, 1, 0, 0, 0, 0]$

The X matrix that represent the observables is constructed with the formula:

$$y_i = \sum_{j=1}^{N} \beta_j X_{i,j} + \epsilon_i$$

where the covariance matrix of the X is fixed, with ones on the main diagonal and ρ in the other entries.

The results of different statistical approach, such as LASSO, RIDGE, PCA and Step Forward selection, to evaluate the MSE changing the parameter are reported below, including some brief explainetion of the theory behind a model reported in each caption of the figures:

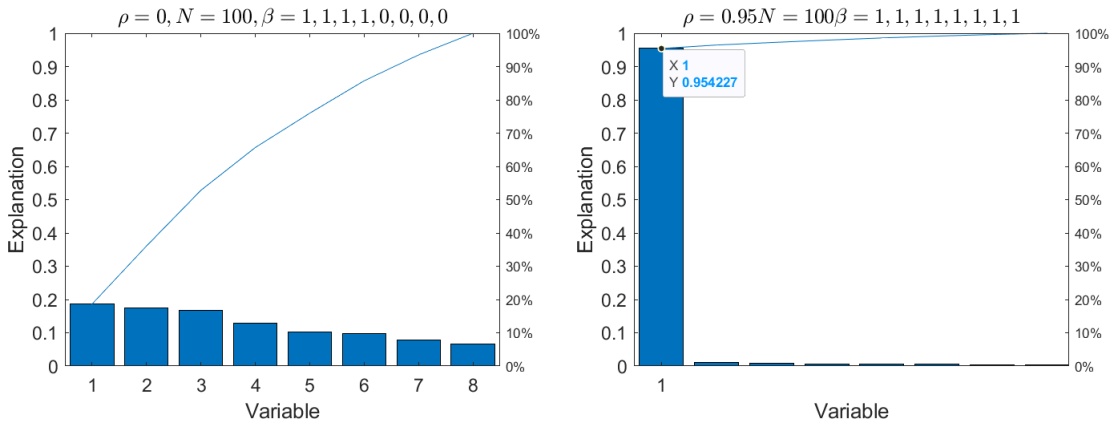


Figure 3: Explanation percentage of predictors with PCA

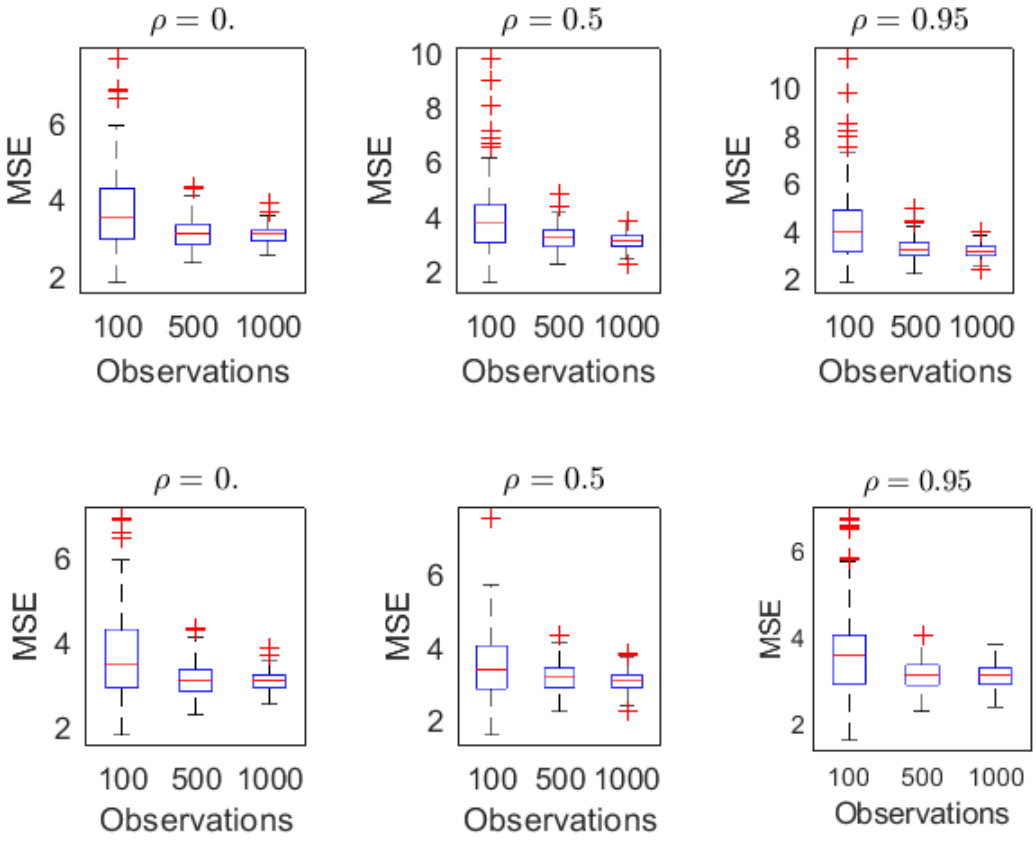


Figure 2: PCA is a statistical method used for dimensionality reduction and data exploration. It transforms a dataset with potentially high-dimensional variables into a new set of uncorrelated variables called principal components. These components are linear combinations of the original variables and capture the maximum amount of variance in the data. PCA helps identify the most informative patterns and reduces the dataset's complexity by identifying the directions of greatest variation

PCA	$\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
DENSO	3.0073	3.1203	3.1144
SPARSO	3.0012	3.0824	3.0562

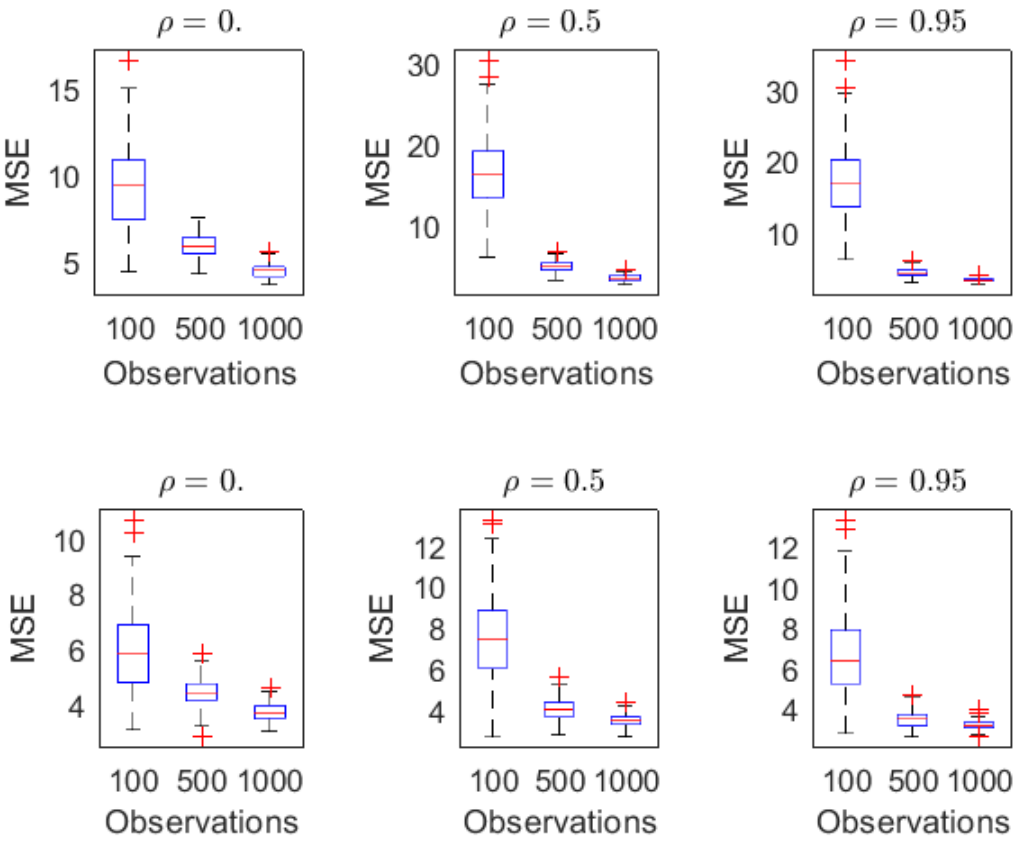


Figure 4: RIDGE regression is a statistical method used for regression analysis that adds a penalty term to the linear regression objective function. The penalty term, determined by a tuning parameter (λ), controls the amount of regularization applied to the regression coefficients. Ridge regression is particularly useful when dealing with multicollinearity as it can mitigate its effects by shrinking the coefficients towards zero without excluding any predictors from the model

RIDGE	$\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
DENSO	4.4455	3.7002	3.4702
SPARSO	3.7449	3.5430	3.2189

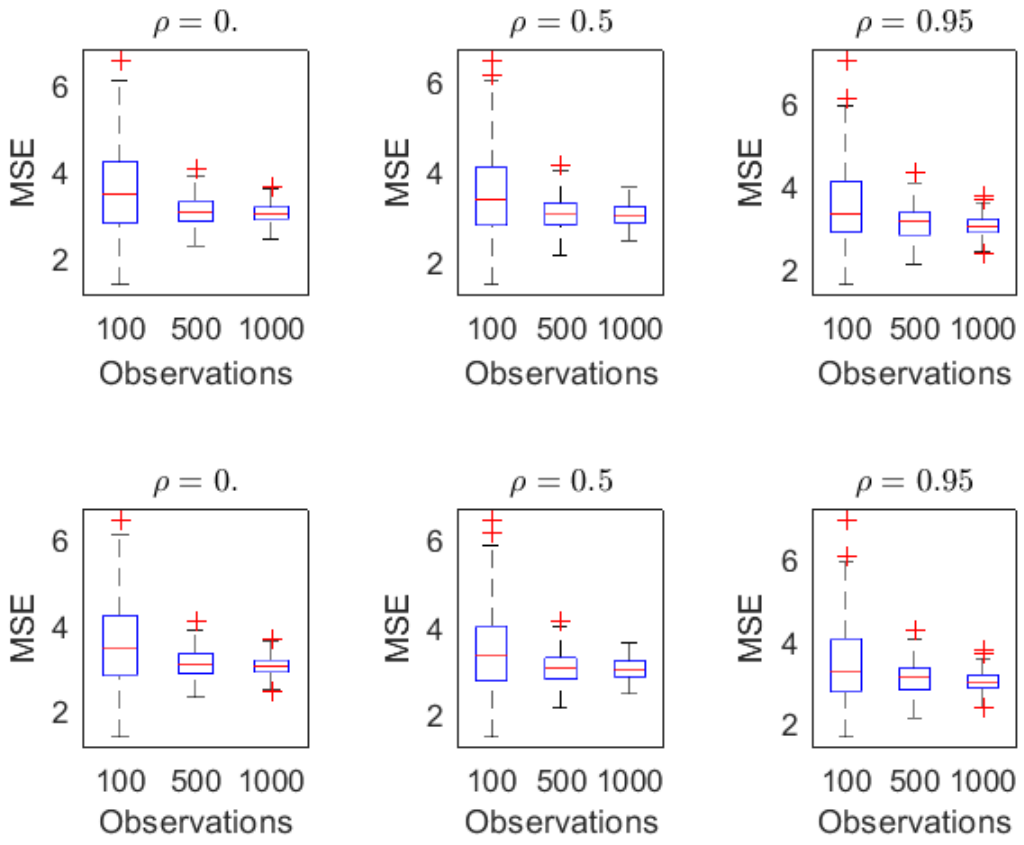


Figure 5: LASSO regression is a statistical method that adds a penalty term to the linear regression objective function. The penalty term, determined by a tuning parameter (λ), encourages sparsity by shrinking some regression coefficients towards zero. This promotes variable selection, effectively excluding irrelevant predictors from the model

LASSO	$\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
DENSO	3.0389	3.025	3.0276
SPARSO	3.04	3.036	3.021

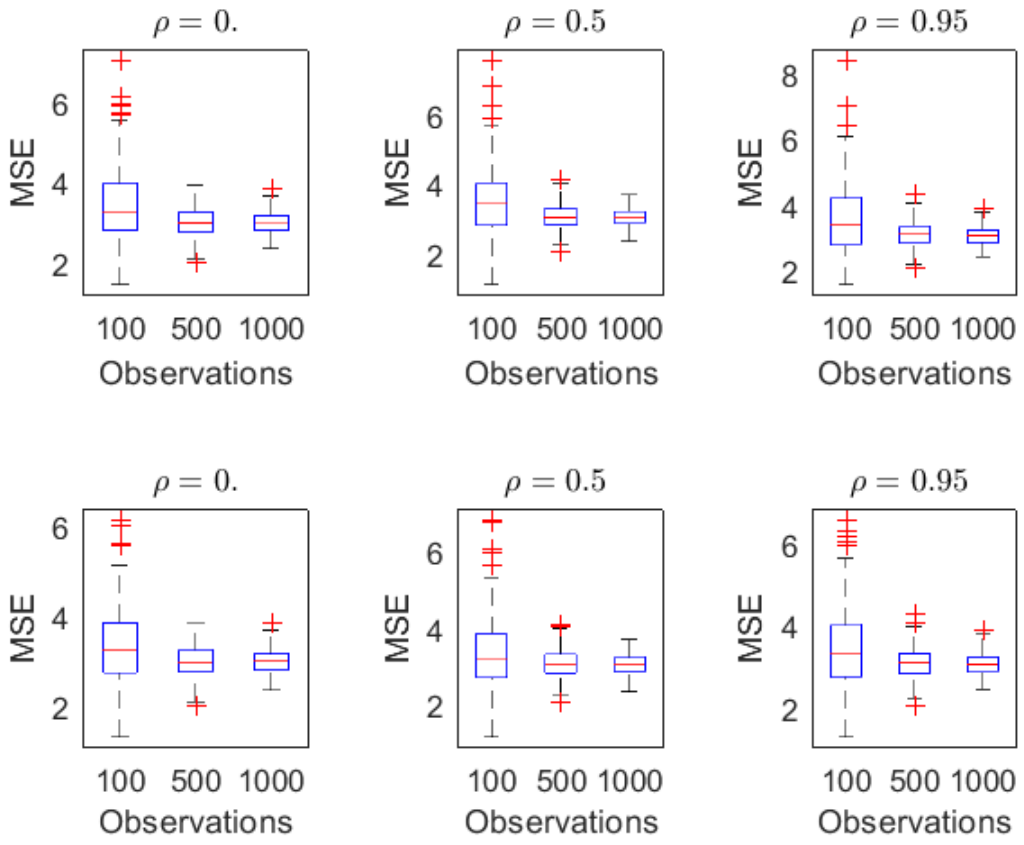


Figure 6: Stepwise forward selection is a statistical model building method used for variable selection in regression analysis. It starts with an empty model and iteratively adds the predictor variable that improves the model's performance the most, based on a specified criterion (e.g., adjusted R-squared or AIC). The process continues until no further improvement is achieved. This method sequentially includes variables in the model and evaluates their contribution, making it useful for identifying the most significant predictors

FW	$\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
DENSO	3.0290	3.0182	3.0540
SPARSO	3.0222	3.0109	3.0413

2.2 Results

Analyzing the MSE for each simulation, is possible to conclude that the best Statistical Model to predict a Sparse dataset is LASSO regression, according to the theory, that reach the expected variance given as input ($\sigma = \sqrt{3}$) with an error of just 0.02, evaluating the set with $\rho = 0.95$. At the same time, this Model is also able to handle with the Sparse Model, with just few predictors relevant!

For Dense Model, is interesting to observe that the best Model that go closer to the expected variance is PCA; this was predictable for its capacity to reduce the space and the complexity in models with a lot of non relevant attributes, and reach this result in the case with $\rho = 0.$, where the variables are not correlated.

Is important to notice that also the Step-Forward selection, that represent the easiest way to avoid extra complexity in easy dataset, like the one in these simulations, achieves very good results for both Models and for every parameter. The RIDGE Model in this case gives the worsts estimations, because it's not allowed to shrink exactly to 0 the parameters, while they are set to 0 in the construction of the dataset!