**Loading the data**

**Process the data**

* We use StringIndexer since it convert categorical features into numerical but keeping its categorical meaning. Vector indexer does the same conversion but gets rid of that data nature information (used when we don’t know the nature of the feature, here we know it is categorical).
* First ufss to divide between categorical and numerical and then conversion to numerical (numerical type but categorical nature remains)

**Creating the model**

**Linear Regression**

Linear regression model was chosen as one of the model for the project because it allows to understand the strength of relationships between the variables. Using R-squared metric it can give clarity on how much total variability in the data is explained by the model. It can as simplify the understanding on which predictors in the model are statistically significant .

After the loading and processing of data steps described above, the vector with following features was created.

|  |
| --- |
| DayOfWeek, CRSElapsedTime, DepDelay, Distance, TaxiOut, DepTime\_index, CRSDepTime\_index, CRSArrTime\_index, Origin\_inde, Dest\_index |

To train the model the regressor takes in the input of features (the vector decribed above) and the name of the colun that represents the feature to be predicted, in this case '*ArrDelay*'. The training is performed on train\_data which includes 7/10 of data.

Following, the model was applied to the test\_data set. The prediction calculated can be found in "Valdating model" section.

**Validating the model**

**Linear Regression**

The metrics metricsof the model found after performing the training are presented in the screenshot below.

Text

Description automatically generated

* **R-squared** measures the strength of the relationship between the input vector of features (*’features*’) and the dependent variable(*’ArrDelay’*). In other words it answers the question: "How much of the data does this model explain? ". The value of 0.939473 means that 93% of the data represent the variance of the dependent variable (*’ArrDelay’*).
* **Root Mean Square Error (RMSE)** is the square root of the average of the squared difference of the predicted and actual value. It is used as ameasure of accuracy, to compare prediction errors of different models. The model with the lowest RMSE is the best one.
* **Coefficience** decribes relationship between the dependent variable (x in our case *vector*) and independent (y in our case ‘*ArrDelay*’) variable. Results presented below show that the attributes with the most significant influence (0ver 0.9) on *ArrDelay* are *DepDelay* are *TaxiOut*. The strength of relationship between *CRSElapsedTime* and our independent varible is 0.2.

Text

Description automatically generated

Prediction metrics compare the predicted value by the model with the actual value of ArrDelay of the testing set.

Table

Description automatically generated

**CrossValidation for Linear Regression**

CrossValidation allows us to compare different machine learning models. Rather the spliting the data into train and test sets crossvalidation separates data into the f*olds* or "blocks" and uses one at the time,and summarizes the results obtained in the end.

For the performing cross validation for linear Regression the amount of 5 folds was chosen. The metrics of the best model found after performing the validation are presented in the screenshot below.

Graphical user interface, text, application

Description automatically generated

The metrics from CrossValidation demonstrate metrics closed to ones from the Linear regression with train/test set . The the same explanation apply in this case.

**Coefficience** as explained above, decribes relationship between the dependent variable. Results presented below show the significance compared with Linear regration with train/testresults, the attributes with the most significant influence (0ver 0.9) on *ArrDelay* are *DepDelay* are *TaxiOut.* The strength of relationship between *CRSElapsedTime* and our independent varible is -0.2, this time negative.

Text

Description automatically generated

**Advanced creterias fullfilled**

* 3 machine learning algorithm
* CrossValidation
* Proper exploratory data analysis
* Smart use of the Spark tools
* Smart handling of special format