Machine learning Project – Milestone 2 report

* Text

  Description automatically generatedReading Data
* Text

  Description automatically generatedAdjusting labels
* Data Engineering and Adjusting Data

Text

Description automatically generated

Text

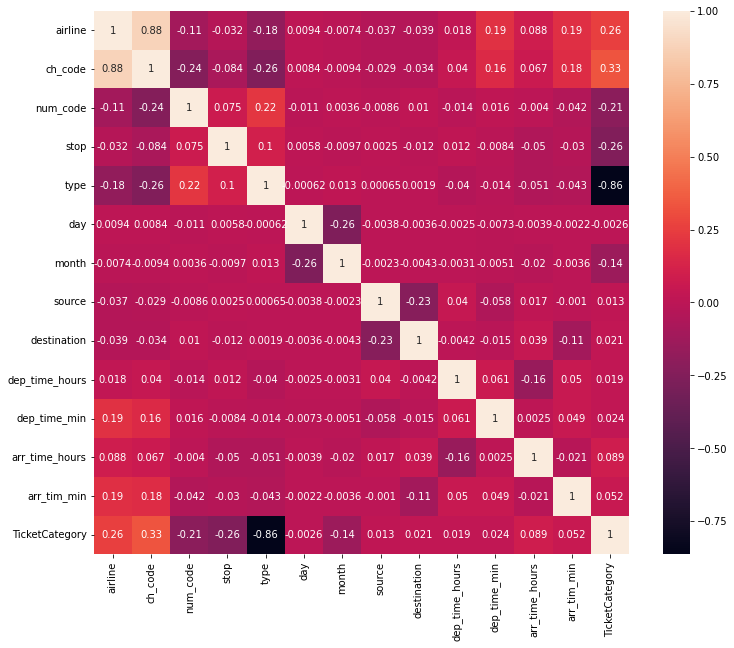
Description automatically generated

* Text

  Description automatically generatedSplitting data and Preprocessing
* Correlation and Feature Selection

Text

Description automatically generated



Note: top\_feature wasn’t used because the model performance was much worse after reducing features.

* Preprocessing Cont’d

Text

Description automatically generated

* Didn’t use top\_feature

Text

Description automatically generatedText

Description automatically generated

Text

Description automatically generated

* LGBMClassifier Model

Text

Description automatically generated

* Performance

Text

Description automatically generated

* XGBClassifier Model

A screenshot of a computer

Description automatically generated with medium confidence

* Performance

Text

Description automatically generated

* Text

  Description automatically generatedRandomForestClassifier
* Performance

Text

Description automatically generated

* Saving all transforming variables

Text

Description automatically generated

* All wanted

Graphical user interface, text

Description automatically generated

* Note: XGB and LGBM training used GPU instead of CPU that’s why they are much faster.
* Random Forest has the best Accuracy
* LGBM is known to be faster than all others but chosen parameters may affected this speed.
* Graph Plot

Chart, bar chart

Description automatically generated

* Conclusion
  + Best Accuracy was RandomForest, but in test samples, RandomForest gave 2 out of 3 unlike both XGB and LGBM were 3 out of 3.
  + Training time and Testing time for RandomForest was bigger than both XGB and LGBM because both XGB and LGBM were trained used GPU unlike RandomForest was trained using CPU.
  + Correlation maps only tells you how much two variables are linearly **correlated**. Which indicates that a direct relationship between them may exist. However, this could not be true. Correlation does not imply causation. That’s why not using correlated features from Heatmap only improved performance.