US airports delays and cancellation

Data Analysis and Classification

Ivan Soto

Introduction to Data Science
ITESM Campus Queretaro

November 25, 2016

Table of Contents

Problem statement

An aviation startup company (Avinalytics) is about to launch a new project, in which they want to understand and characterize the delays and cancellations to create an app capable to determine the likelihood of a flight in the US being cancelled.
First, they want to analyze historical flight data and identify patterns and characteristics for the different flight delays and cancellations, based on the date, the airline operators, the airports, route structures, region or state of the country, etc. The Avinalytics company hires you to perform this task.

Approach to a solution

For every observation in the flights dataset, the outcome of the flight is stated. This allows us to consider this problem as a supervised learning task, studied under the field of Machine Learning. To be more precise, there are two possible outcomes for each flight: cancelled or not cancelled.

Descriptive analysis

Feature selection

Out of the 110 features available for each observation in the data set, a subset of those were selected as representative features through manual selection.

- Quarter
- Month
- UniqueCarrier
- FlightNum
- OriginAirportID
- DestAirportID
- CRSDepTime
- DepDelay
- CRSArrTime
- ArrDelay
- Cancelled
- CancellationCode
- Distance

Descriptive analysis

Categorical features

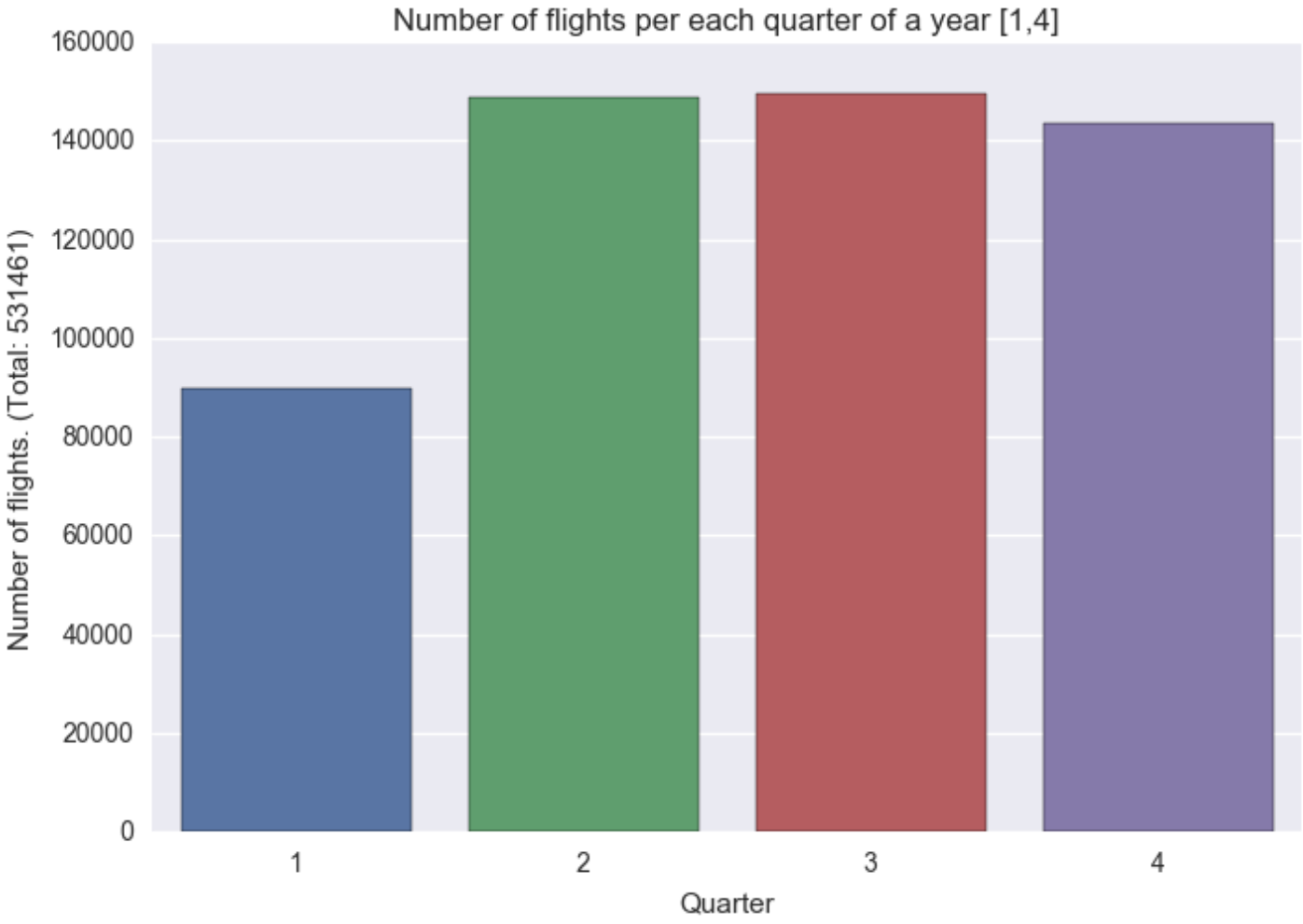


Figure: Flights per year quarter

Descriptive analysis

Categorical features

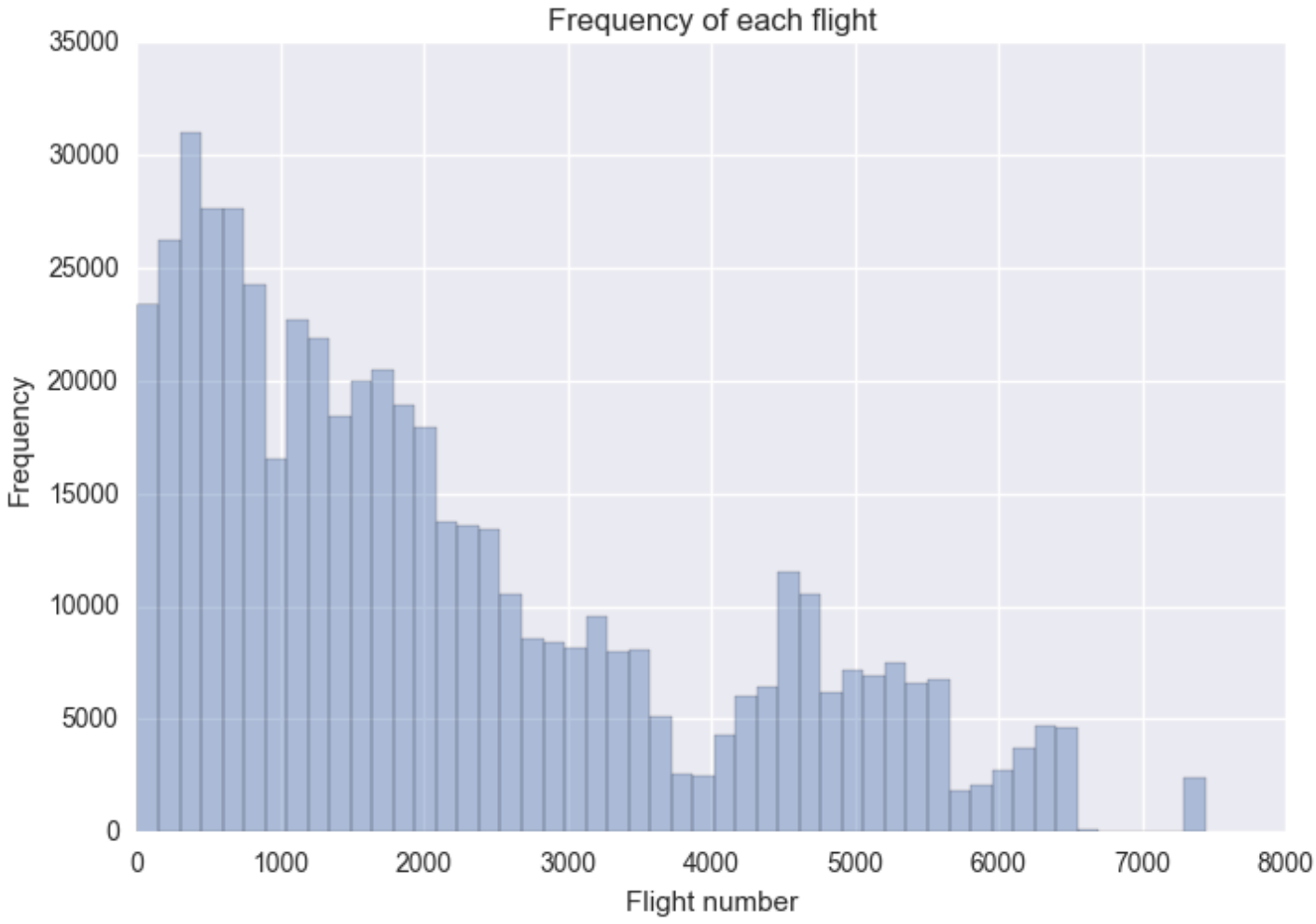


Figure: Frequency of flights

Descriptive analysis

Non-categorical features

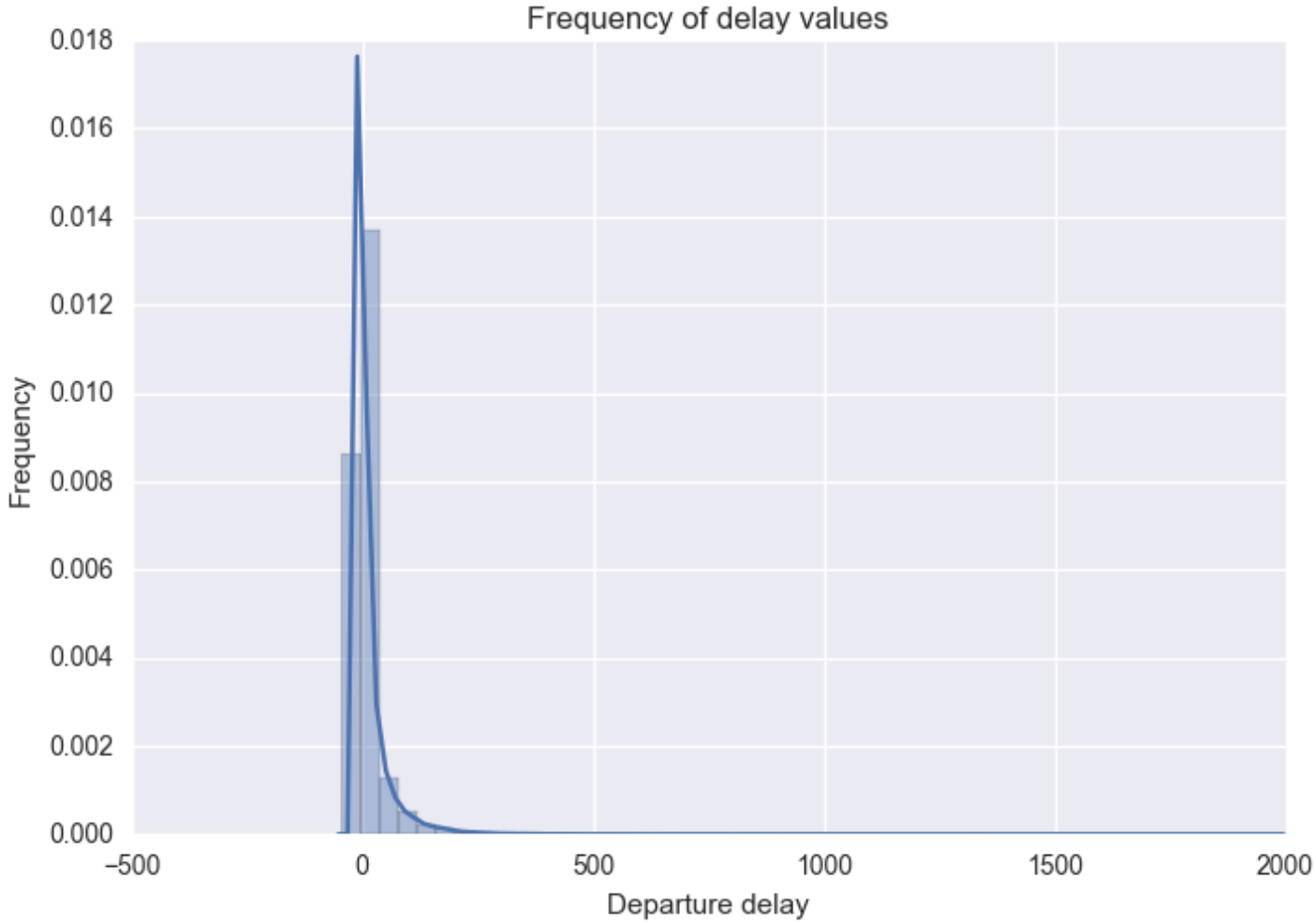


Figure: Delay on flights

Descriptive analysis

Non-categorical features

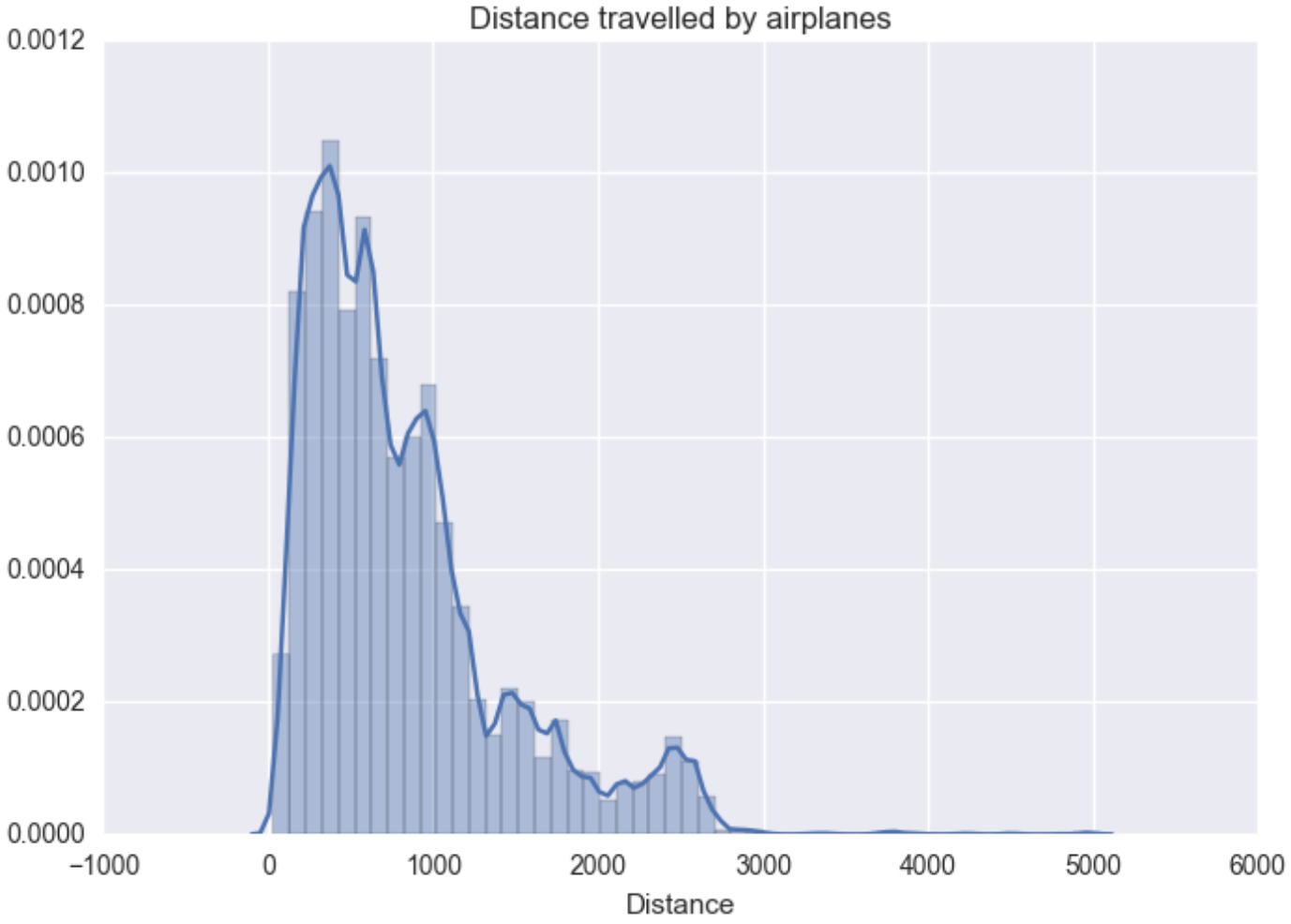


Figure: Distance of flights

Descriptive analysis

Airports

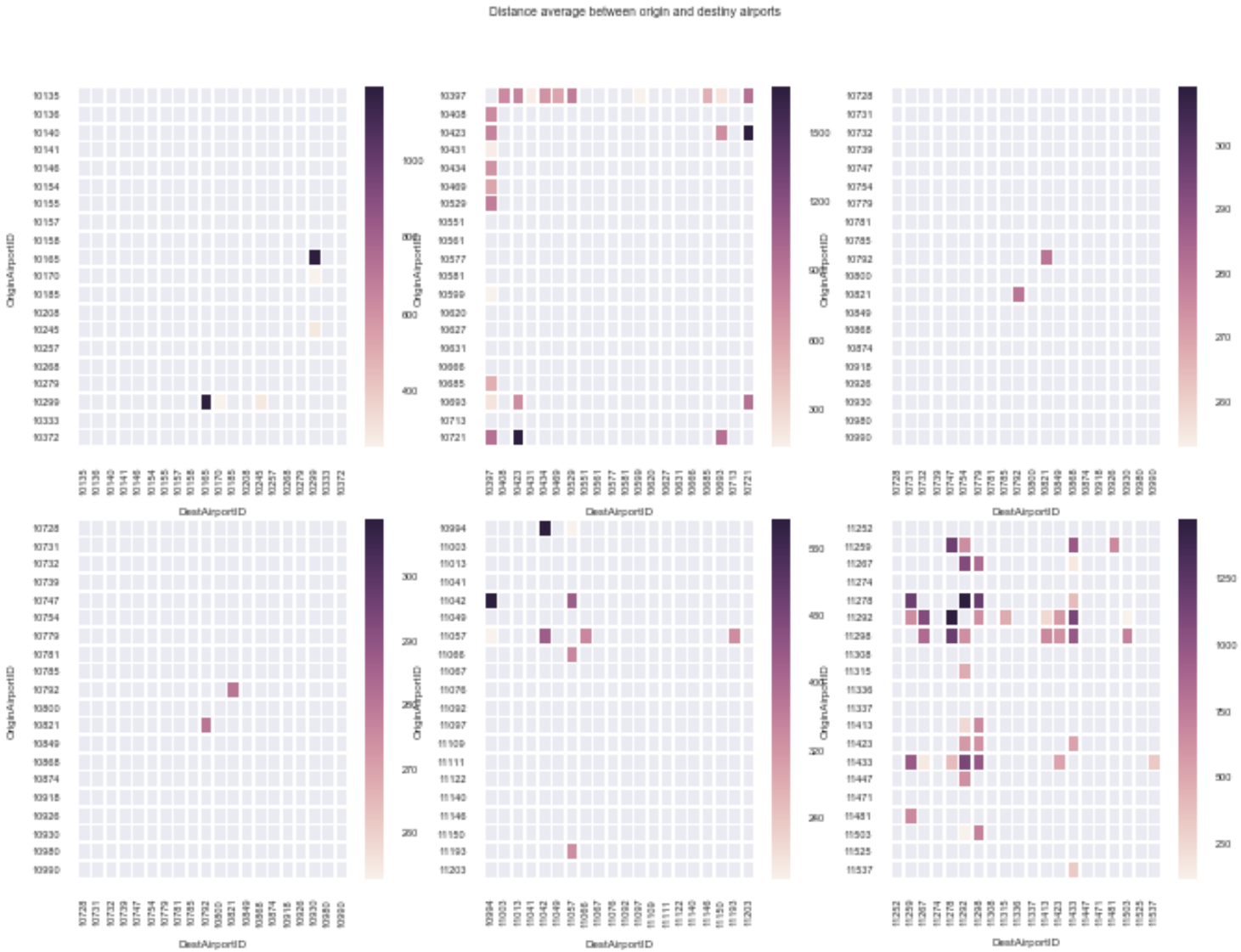


Figure: Mean distance between airports

Descriptive analysis

How many are cancelled?

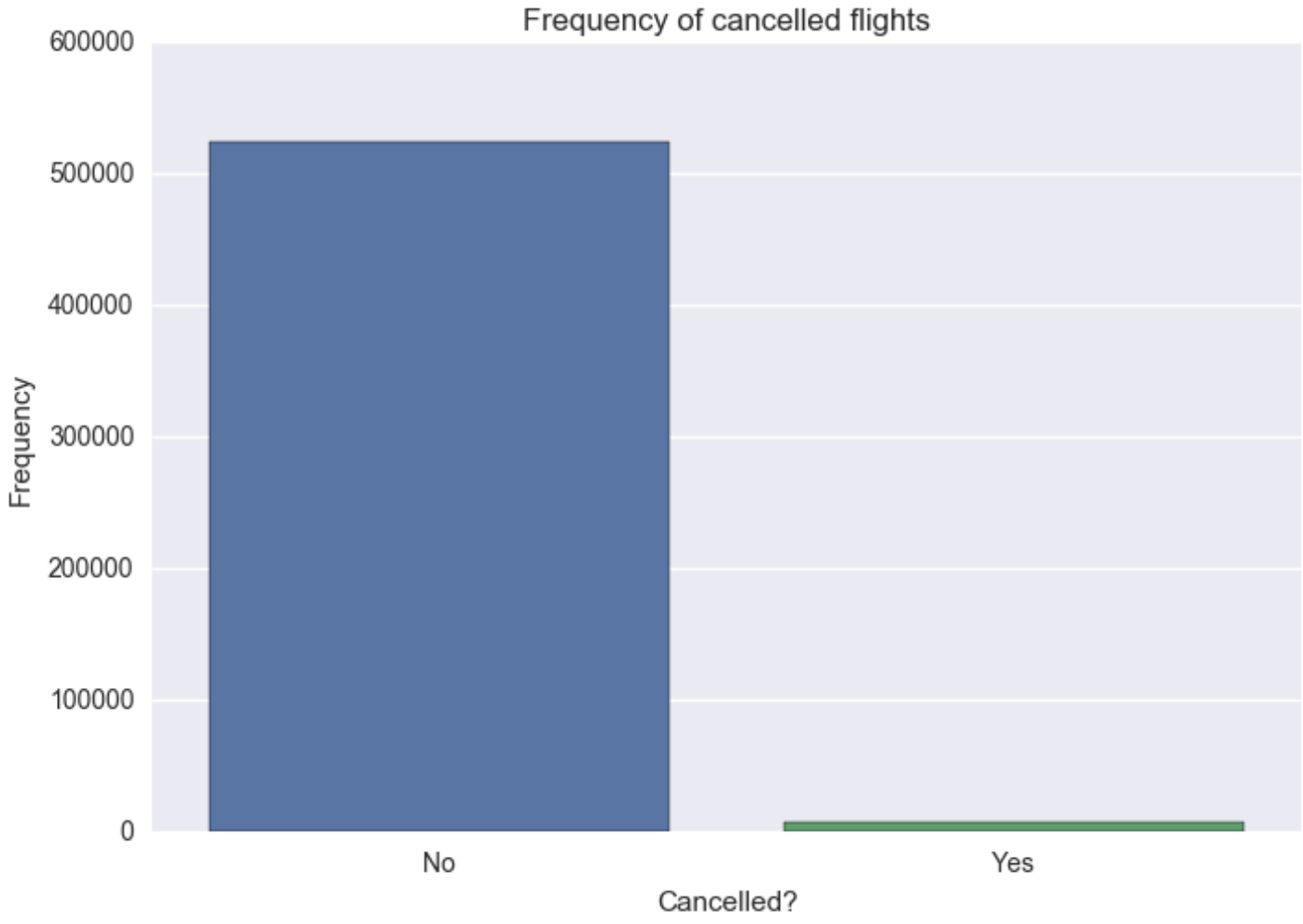


Figure: Cancelled flights

Exploratory analysis

Correlations

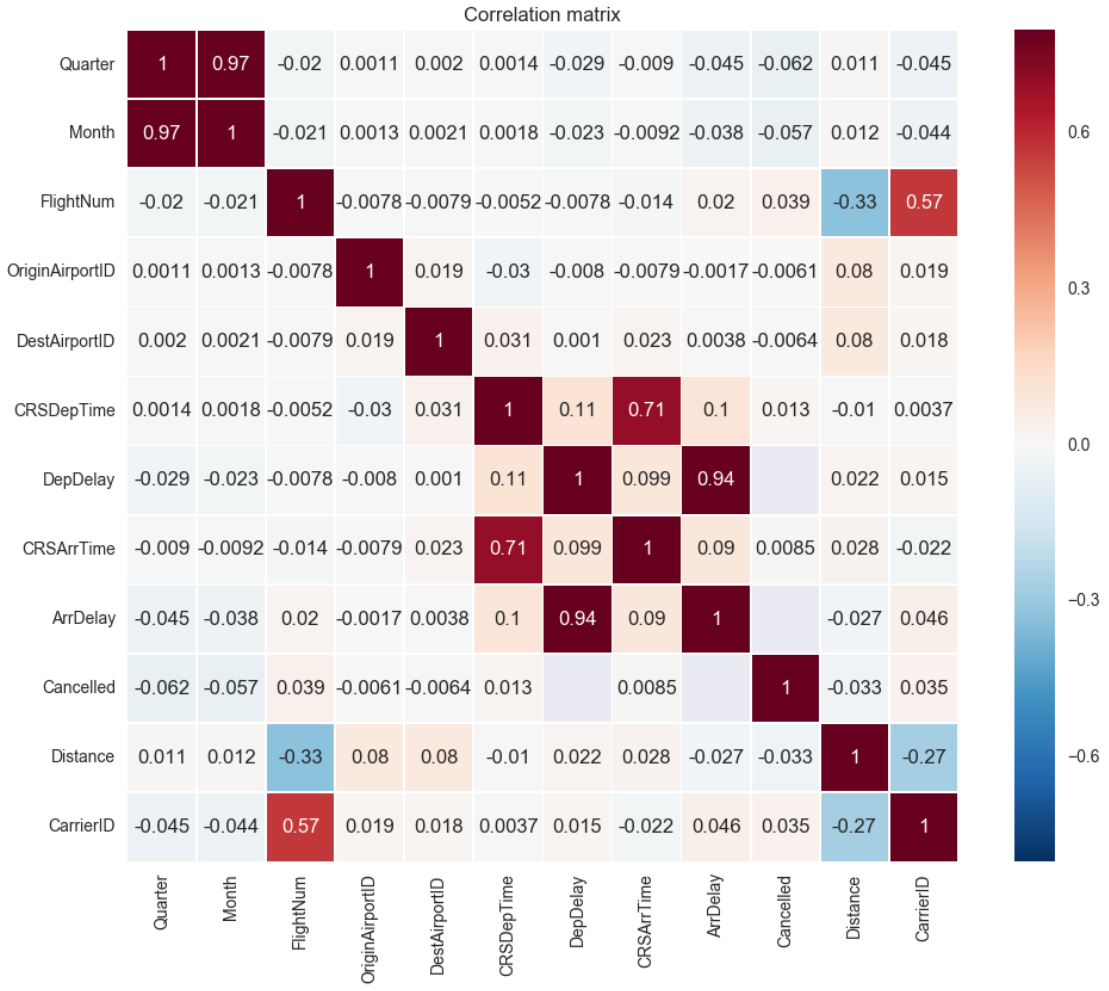
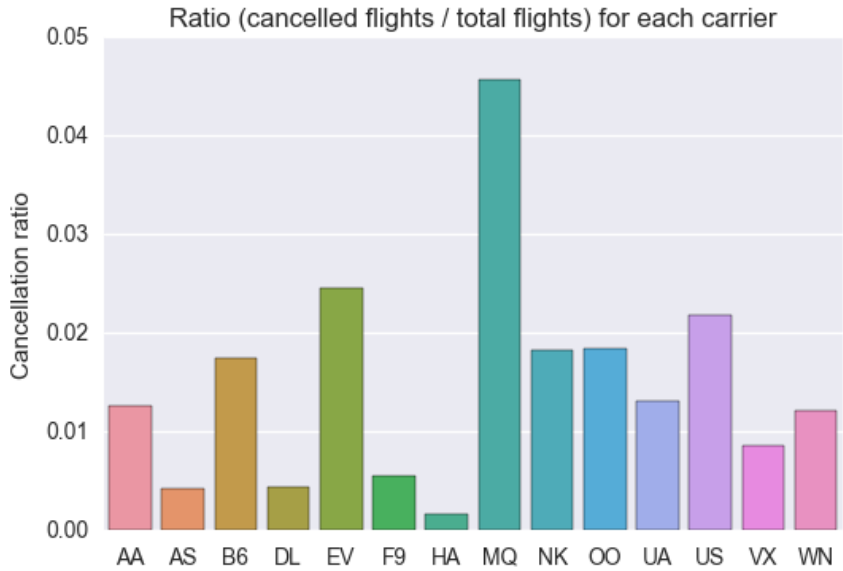
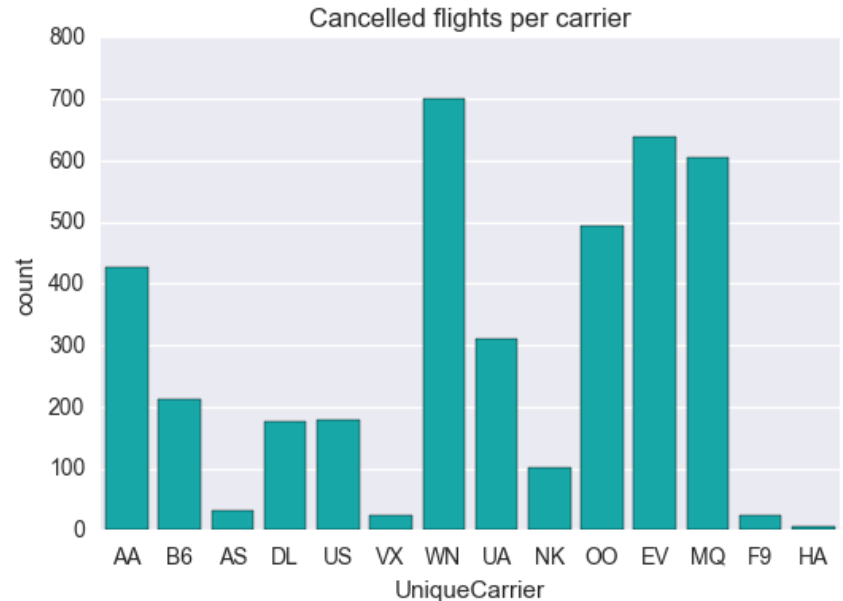


Figure: Heatmap on correlation matrix

Exploratory analysis

Correlations





Exploratory analysis

Correlations

Figure: Scatterplot showing each variable against the others

Exploratory analysis

Correlations

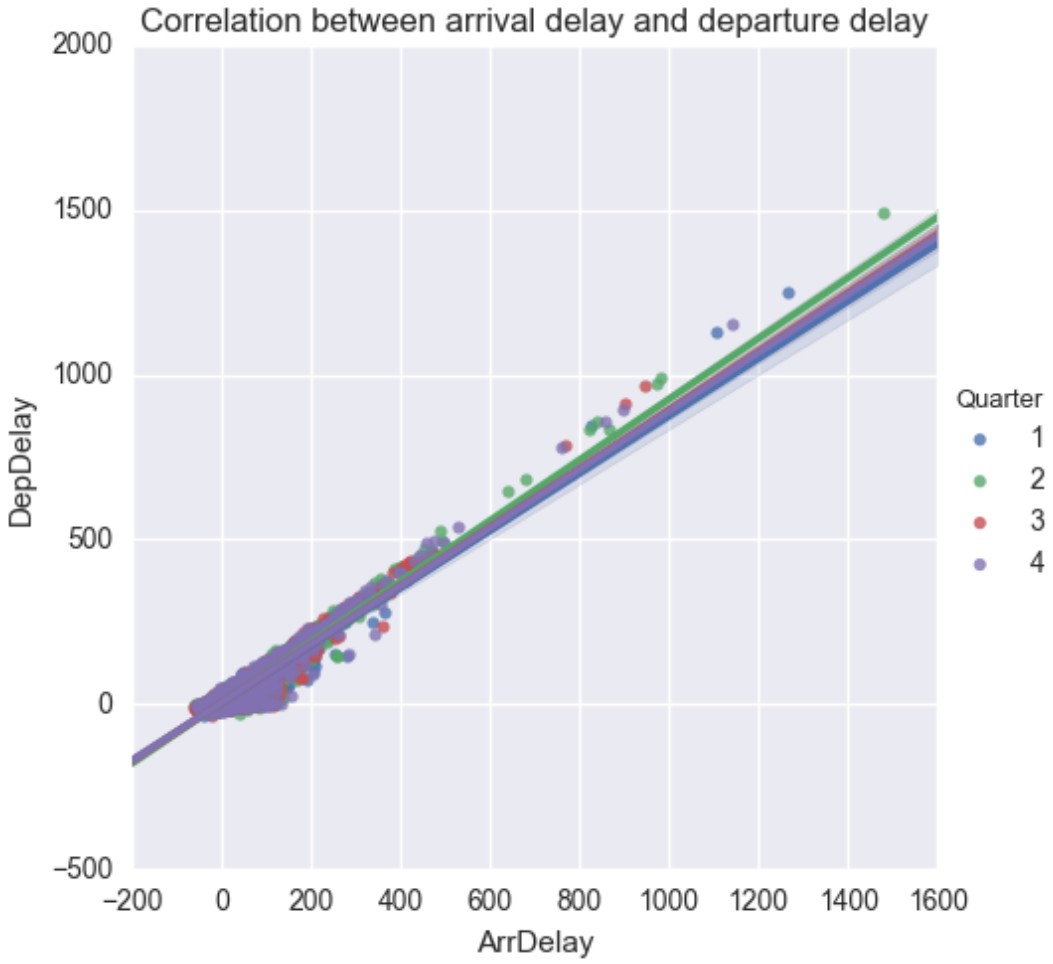


Figure: Regression between arrival delay and departure delay

Exploratory analysis

Correlations

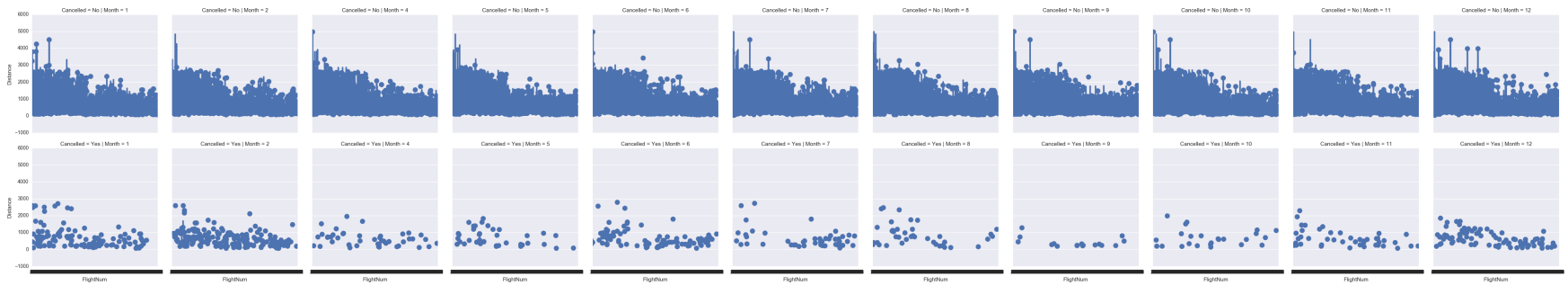


Figure: Distances for cancelled and not cancelled flights

Classification

Dividing the data

Several classifiers were trained using 85% of the 40% sample obtained out of the whole dataset. The reason is that, a classifier should never be trained using the complete dataset: it's going to fail at predicting unseen data (i.e. it's going to be overfitted).

Classification

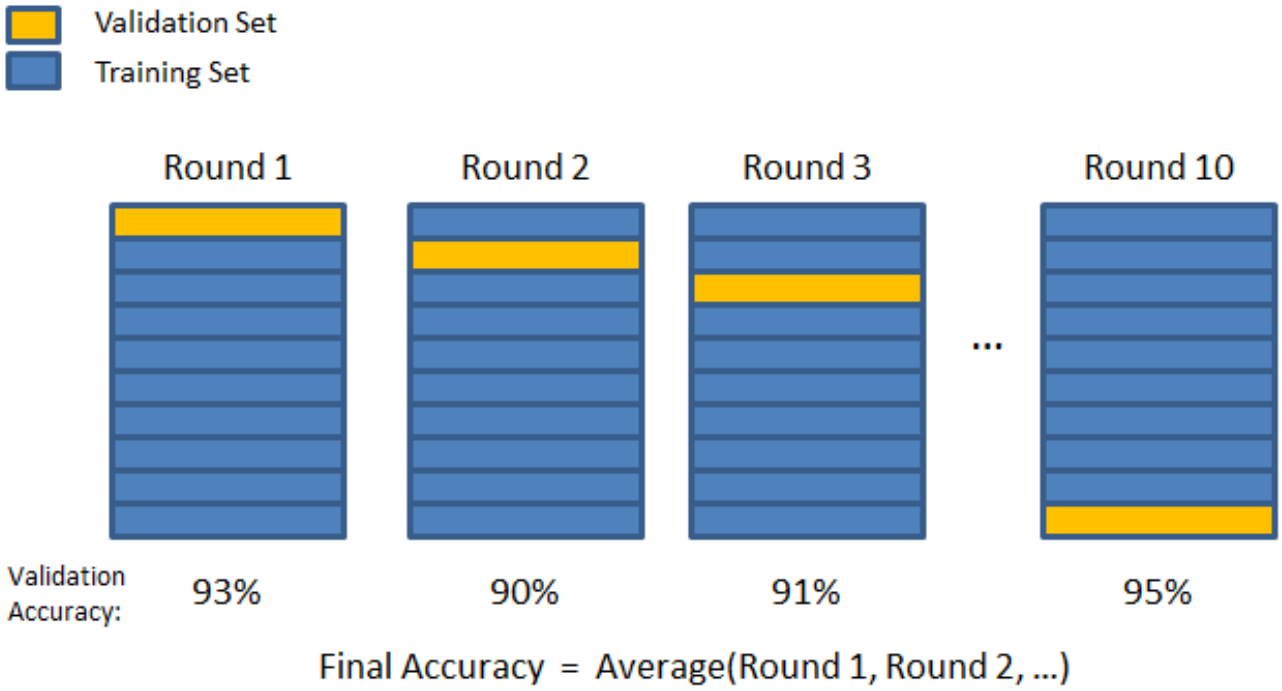
Searching for the best hyperparameters

Hyperparameters are not directly learnt from the classifier. Instead, they are set and tuning them according to the problem at hand can improve the results drastically. Depending on the classifier, the hyperparameters that are available for tuning may vary, and there are several approaches to finding the best hyperparameters. The most used algorithms are known as Grid Search and Randomized Search.

Classification

Cross-validation

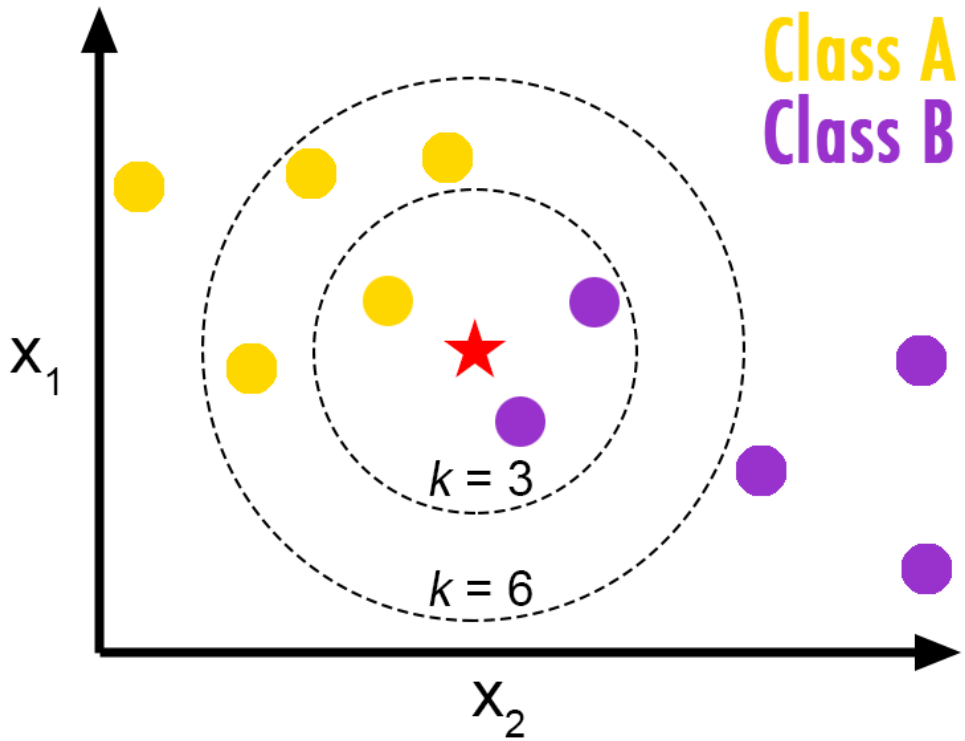
A cross-validated search for the best parameters is performed on both Grid Search and Randomized Search, thus avoiding the problem of overfitting. Cross-validation (also known as k-fold cross-validation) is a technique that divides the dataset into k pieces. $k-1$ pieces are used to train the classifier, and the remaining kth piece is used to test the model.



Classification

K-nearest neighbors

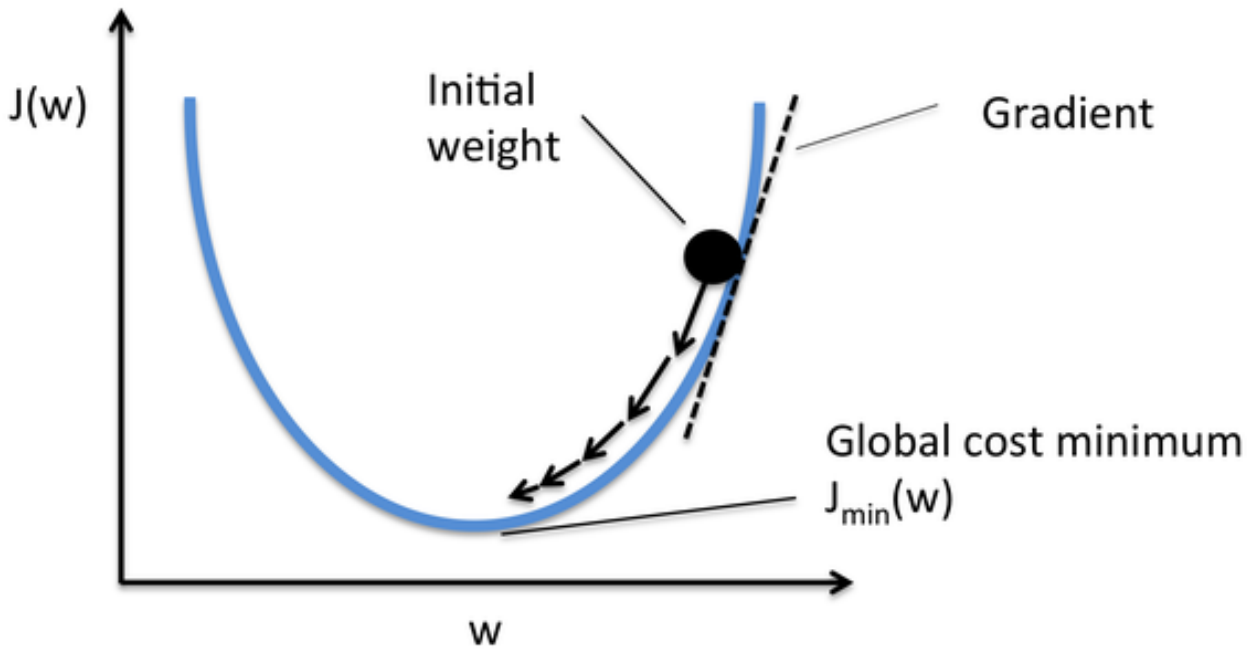
An observation for which we want to predict a class can be vectorized in $\mathbb{R}^n$ space. k votes decide the class this trial belongs to.



Classification

Stochastic Gradient Descent

Gradient Descent is an optimization algorithm to find the local minima in a function $f(x)$. In the context of supervised learning, Gradient Descent is used to find the optimal coefficients of a vector $\vec{v}$ that minimizes a cost/loss function.

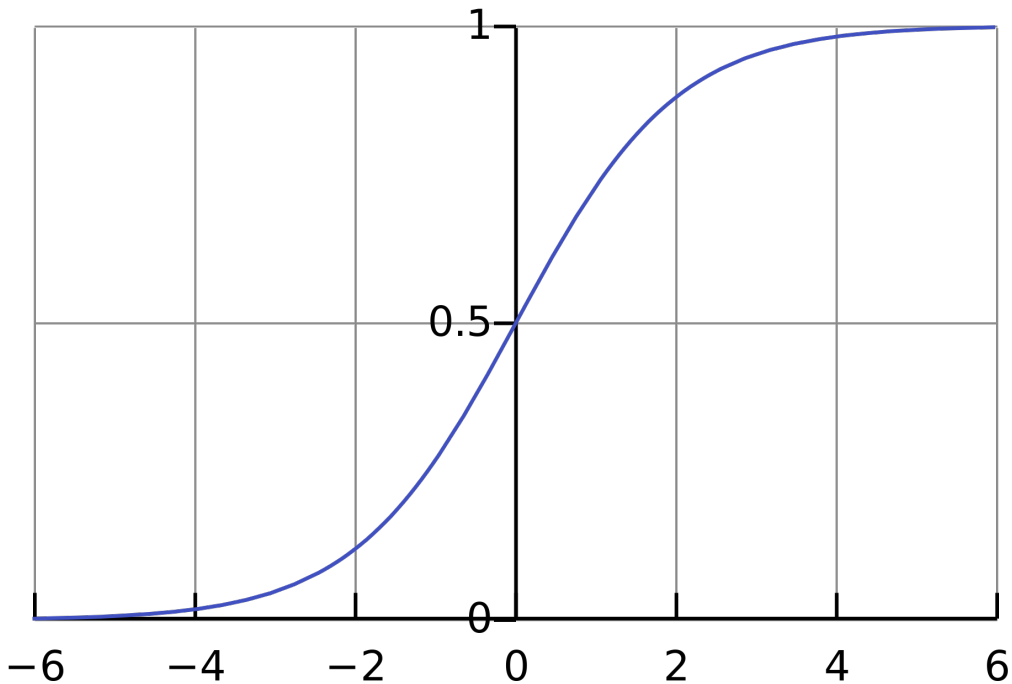


Classification

Logistic regression

For each trial the probability that it belongs to a certain class is calculated using a logistic function (represented as $f(x) = \frac{1}{1+e^{-x}}$). L1 and L2 regularization can be used.

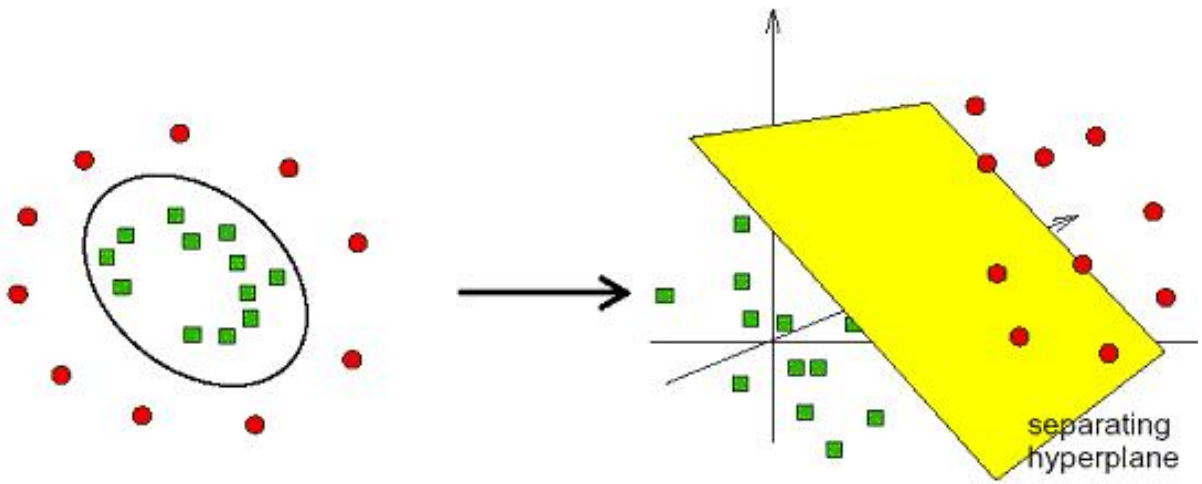
$$min_f \sum_{i=1}^{n} V(f(\hat{x}_i), \hat{y}_i) + \lambda R(f)$$



Classification

Support Vector Machine

A SVM (Support Vector Machine) builds a hyperplane $\vec{w}$ between the points so that it becomes a decision boundary that divides points into two classes. A trial is classified based on the side of the decision boundary it is in.



Classification

Naive Bayes

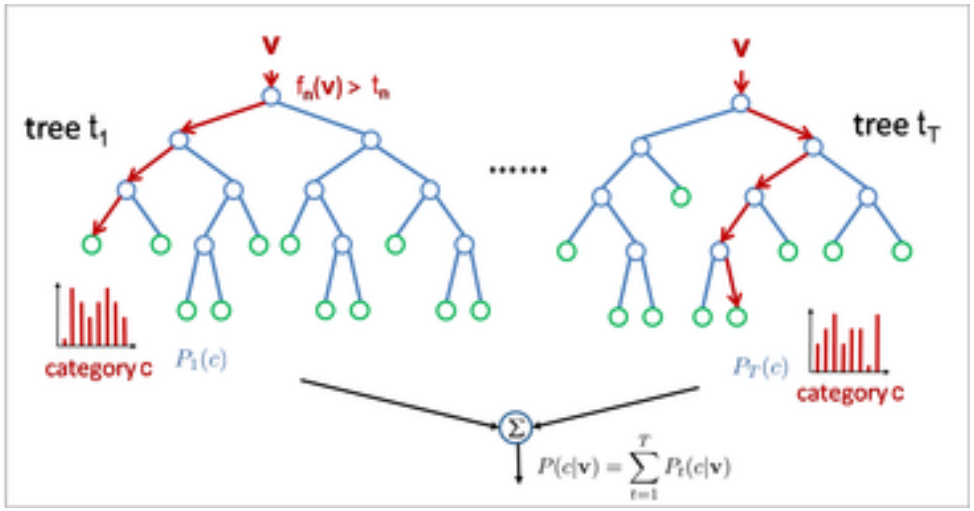
The idea behind the series of algorithms under "Naive Bayes methods" is to apply Bayes' theorem with a naive assumption that every pair of features are independent. Bayes' theorem is represented by the equation

$$P(y \mid X) = \frac{P(X \mid y)\, P(y)}{P(X)}$$

Classification

Random Forest

Random Forest consists of several decision trees, each independent to every other, and every tree is learns from a random sample drawn with replacement, and at the end, the mode of the results (for the classifier version) is taken as the predicted class.



Classification

Voting classifier

Last but not least, another ensemble was built. Not out of decision trees this time, but out of the Logistic Regression, Random Forest and Bernoulli Naive Bayes classifiers. This is also known as a voting classifier: each classifier predicts a class for the trial, and soft voting took place to make a decision on the final class.

Classification

Visualization for accuracy scores

Figure: Accuracy scores for the trained classifiers

Classification

Visualization for F1 scores

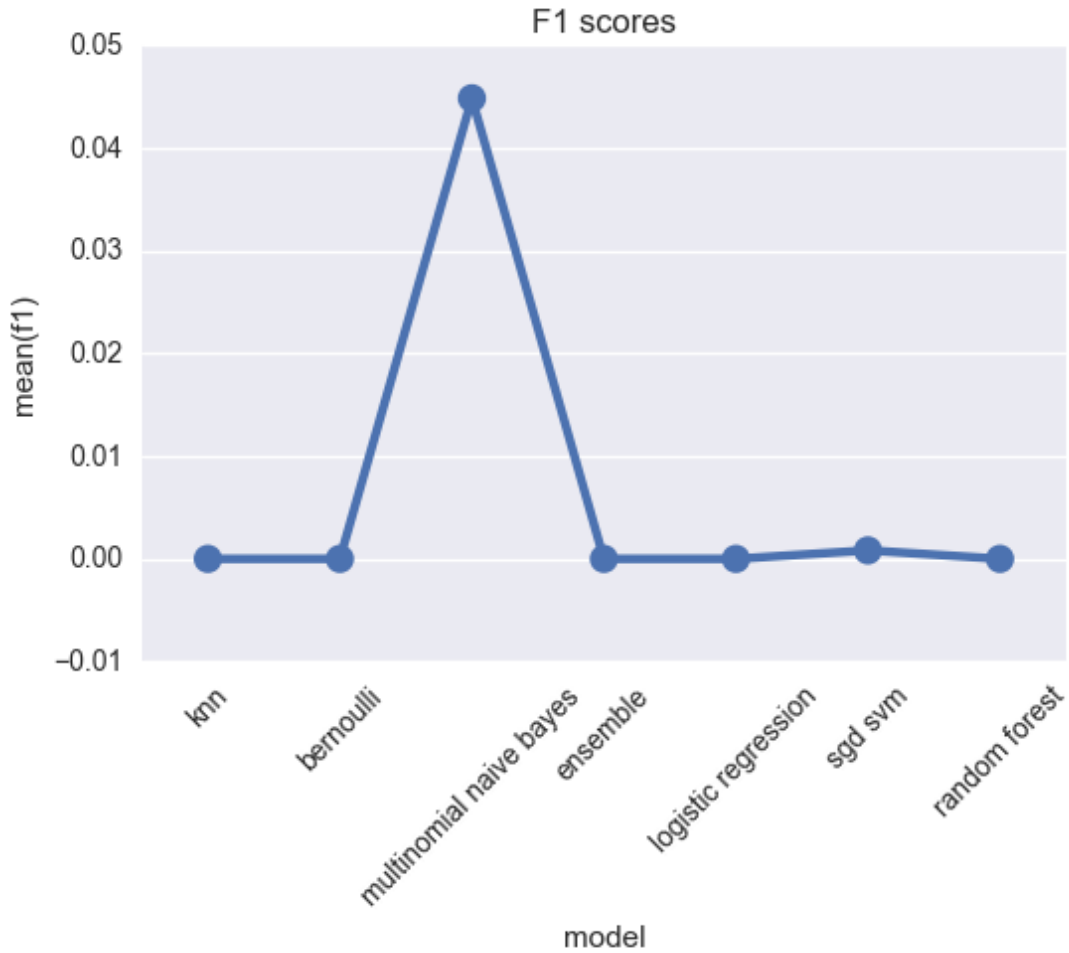


Figure: F1 scores for the trained classifiers

Conclusion

Appendix and references

The code for the descriptive, exploratory analysis and classification task can be found in this repository, along with the serialized models:
https://github.com/IvanAli/DataScienceITESM

- On-time flights (2016). Document retrieved from: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
- Flight delay cause (2016). Document retrieved from: http://www.transtats.bts.gov/ot_delay/ot_delaycause1.asp?type=21&pn=1
- Flight Stats Global Cancellation and delays (2016). Document retrieved from: http://www.flightstats.com/go/Media/stats.do
- Supervised learning – scikit-learn 0.18.1 documentation. Document retrieved from: http://scikit-learn.org/stable/supervised_learning.html
- Ensemble methods – scikit-learn 0.18.1 documentation. Document retrieved from: http://scikit-learn.org/stable/modules/ensemble.html