|  |
| --- |
| UC Santa Scruz Silicion Valley extension |
| Predict airline delay |
| Final project – Python for Programmers |
|  |
| **Dimple Sharma** |
| **3/25/2015** |

|  |
| --- |
| Final project of Python for programmers class. The project implements data analysis on airline dataset. The goal is to predict flight delay using supervised learning model. |

Contents

[Introduction 3](#_Toc415067088)

[Requirements 3](#_Toc415067089)

[Dataset 3](#_Toc415067090)

[Description 5](#_Toc415067091)

[Import Modules 5](#_Toc415067092)

[Read data 5](#_Toc415067093)

[Data exploration 5](#_Toc415067094)

[Data preprocessing and feature matrix 7](#_Toc415067095)

[Analyze and build model 8](#_Toc415067096)

[Screenshots 8](#_Toc415067097)

[Conclusion 8](#_Toc415067098)

[Python Program 9](#_Toc415067099)

[References 9](#_Toc415067100)

# Introduction

Approximately 20% of airline flights are cancelled or delayed each year which results in significant cost for both travelers and airlines. This project aims to build a predictive model which will be supervised learning model to predict airline delay using historic flight data.

The goal is to perform step by step data analysis for a supervised learning problem using airline dataset. This project provides solution for building supervised learning algorithm such as logistic regression. It explores the airline dataset using dataframe object from pandas module. It implements data analysis techniques of data cleaning, data transformation, data exploration, data analysis, and visualizing data.

This project provides solution for using python and matplotlib to explore raw dataset. It uses attributes and functions of data frame in pandas module to transform raw data into feature matrix. It uses Scikit-learn machine learning library for building a predictive model.

Since airline dataset is large and has many rows, to simplify we build a supervised learning model to predict flight delays for flights leaving O’Hare International airport (ORD).

# Requirements

This implementation uses open source modules for Python therefore it is easy to reproduce. The modules can be installed using familiar installation tools like pip.

In this project we use the python libraries :

* Pandas
* Numpy
* Scikit-learn
* Matplotlib
* Sys
* Warnings

# Dataset

This project uses airline dataset which is available at <http://stat-computing.org/dataexpo/2009/the-data.html> The website provides airline dataset about flights in the US from the years 1987-2008. However, in this project we use airline dataset for year 2007 and 2008 only. The airline dataset contains the 29 variables described as follows:

|  |  |  |
| --- | --- | --- |
|  | Name | Description |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | unique carrier code |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | Origin |
| 18 | Dest | destination |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

# Description

In this project we will build a supervised learning model to predict flight delays for flights leaving O'Hare International airport (ORD). We use supervised learning model to learn the model using historic data from 2007 and evaluate its performance using data from 2008.

To build a predictive model for predicting airline delay, the target variable will be departure delay which has variable name “DepDelay” and it provides flight delays in minutes. We build a classification model to predict flight delay and to achieve this we transform our target variable into a binary variable by defining a "delay" as having 15 mins or more of delay, and "non-delay" otherwise. The response variable for training set “train\_y” is a binary variable, 0 meaning flight not-delayed and 1 meaning flight is delayed. Similarly the response variable for test set “test\_y” is also a binary variable.

## Import Modules

We import the modules that we will need for data analysis. We import sklearn module which provide key function for linear algorithm. We import pandas module which provides key functions to read csv file and work with data frame objects. We import matplotlib which provides functions to create data visuals.

## Read data

The airline dataset is a comma separated file, the values in a row are separated by comma and each row is separated by new line. The data is read using read\_csv function in pandas which returns the data in a data frame object.

## Data exploration

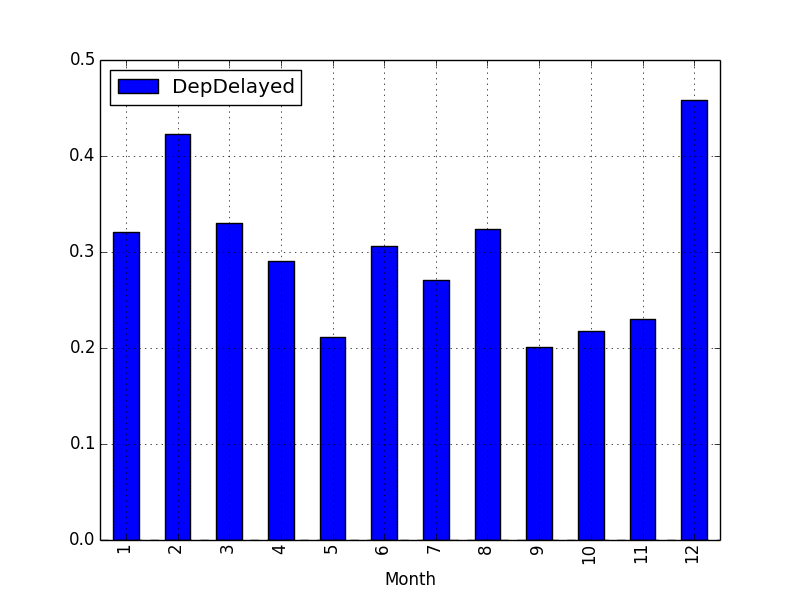
Data exploration is the first and most common step in building predictive model. We begin with exploring the data first to get an understanding of the data and get clues to which features might be good for predictive modeling.

After reading the data, we provide a description of the dimensions and columns to learn more about the data. We see there are 7.4M+ flights in 2007 and 29 variables. We further explore data to look at basic statistics after limiting the data to O’Hare airport. We get following output:

total flights: 359169

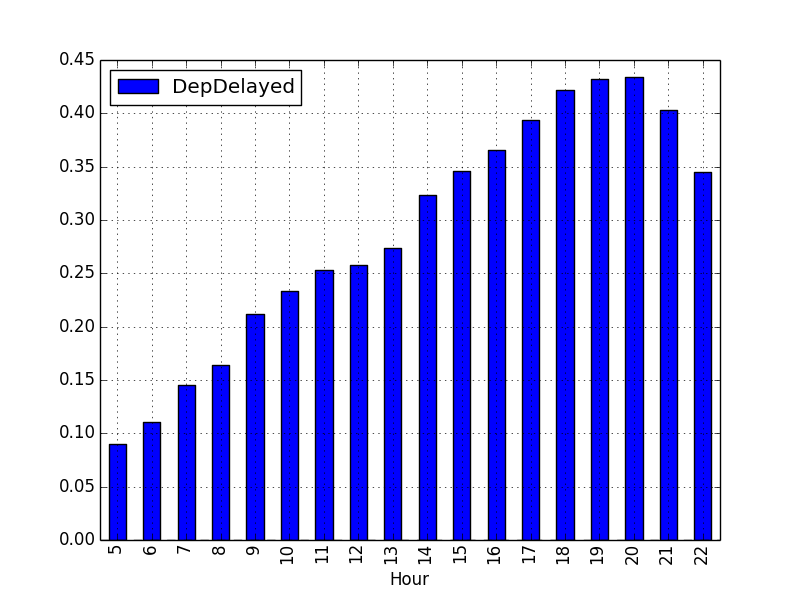
total delays: 109346

Following is an example of delayed flights distributed by months:



We see highest number of delays in December and February, December is a holiday season and February is the peak of winter.

Following is an example of flights delayed by hour of day:



We see that flights tend to be delayed later during the day.

## Data preprocessing and feature matrix

Data exploration provided insights into our dataset so the next step will be to preprocess the data to build feature matrix. We use the variables in our dataset to identify predictor variables for our model.

In this step we will build a feature matrix for our predictive model. This step often requires data transformation in which we extract more value from existing variable.

The following features are identified for our predictive model:

* month: Winter months could have more delays compare to summer months
* day of month: It is likely not a very predictive variable, but we will keep it.
* day of week: Compare delays on weekend vs. weekday
* hour of the day: hours later during the day tend to have more delays
* Carrier: Some carriers can be more prone to delays than others
* Destination airport: Some airports are more prone to delays than others
* Distance: It will be interesting to see distance is a good predictor of delay
* number of days from closest national holiday: we assume that holidays tend to be associated with more delays

We create a util class which generates holidays for year 2007 and 2008 and has a function which calculates days close to holidays.

We use apply function with lambda expression of data frame to extract hour from variable “CRSDepTime”.

In addition, we filter data to keep data for flights that originate from O’Hare airport and were cancelled.

## Analyze and build model

We have built a feature matrix which we will use as predictors to build a predictive model. The next step is to build a binary classification model using logistic regression. We apply preprocess steps and build a feature matrix for airline data from 2007 and 2008. We create training and test dataset using data for year 2007 and 2008 respectively. All predictors are numeric column. Our training set has 359 thousand rows and 6 columns.

We use Scikit-learn machine learning package to build our predictive models using Logistic regression and evaluate its performance using testing dataset. We also create a confusion matrix which counts the true positive, true negatives, false positives and false negatives. Using confusion matrix, we compute precision, recall, F1 metric and accuracy.

# Screenshots

Successfully read airline data.

Data type <class 'pandas.core.frame.DataFrame'>

Number of rows and columns: (7453215, 29)

<class 'pandas.core.frame.DataFrame'>

Int64Index: 7453215 entries, 0 to 7453214

Data columns (total 29 columns):

Year int64

Month int64

DayofMonth int64

DayOfWeek int64

DepTime float64

CRSDepTime int64

ArrTime float64

CRSArrTime int64

UniqueCarrier object

FlightNum int64

TailNum object

ActualElapsedTime float64

CRSElapsedTime float64

AirTime float64

ArrDelay float64

DepDelay float64

Origin object

Dest object

Distance int64

TaxiIn int64

TaxiOut int64

Cancelled int64

CancellationCode object

Diverted int64

CarrierDelay int64

WeatherDelay int64

NASDelay int64

SecurityDelay int64

LateAircraftDelay int64

dtypes: float64(7), int64(17), object(5)

memory usage: 1.7+ GB

None

Year Month DayofMonth DayOfWeek DepTime CRSDepTime ArrTime \

0 2007 1 1 1 1232 1225 1341

1 2007 1 1 1 1918 1905 2043

2 2007 1 1 1 2206 2130 2334

3 2007 1 1 1 1230 1200 1356

4 2007 1 1 1 831 830 957

CRSArrTime UniqueCarrier FlightNum ... TaxiIn TaxiOut \

0 1340 WN 2891 ... 4 11

1 2035 WN 462 ... 5 6

2 2300 WN 1229 ... 6 9

3 1330 WN 1355 ... 3 8

4 1000 WN 2278 ... 3 9

Cancelled CancellationCode Diverted CarrierDelay WeatherDelay NASDelay \

0 0 NaN 0 0 0 0

1 0 NaN 0 0 0 0

2 0 NaN 0 3 0 0

3 0 NaN 0 23 0 0

4 0 NaN 0 0 0 0

SecurityDelay LateAircraftDelay

0 0 0

1 0 0

2 0 31

3 0 3

4 0 0

[5 rows x 29 columns]

['Year' 'Month' 'DayofMonth' 'DayOfWeek' 'DepTime' 'CRSDepTime' 'ArrTime'

'CRSArrTime' 'UniqueCarrier' 'FlightNum' 'TailNum' 'ActualElapsedTime'

'CRSElapsedTime' 'AirTime' 'ArrDelay' 'DepDelay' 'Origin' 'Dest'

'Distance' 'TaxiIn' 'TaxiOut' 'Cancelled' 'CancellationCode' 'Diverted'

'CarrierDelay' 'WeatherDelay' 'NASDelay' 'SecurityDelay'

'LateAircraftDelay']

Total flights:

# Conclusion

Python provides open source modules which can be used for data analysis. The open source modules can be used to build supervised learning models and other kinds of model.

In this project we build a supervised learning model using open source modules Pandas, Sickit-learn, numpy and matplotlib. We built a binary classification model which predicts airline delays. We have used Pandas to perform various types of data pre-processing and feature engineering tasks. We applied Scikit-learn machine learning algorithm on the resulting datasets and evaluated the model using test dataset.

The key output of the program is a python script which helps to build a predictive model. Predictive model is simply a formula which can be used for making predictions. The python script provides text output for data analysis steps which describes dataset dimensions, columns, etc. It also creates 2 graphs using matplotlib.

# Python Program

See attached file

# References

<http://hortonworks.com/blog/data-science-apacheh-hadoop-predicting-airline-delays/>

<http://stat-computing.org/dataexpo/2009/the-data.html>