Suresh Gourigolla

Medicare Admission Analysis

**Introduction:**

**Machine learning helps predict admissions, readmissions. Admission and the number of days in hospital are important for managing the schedule of doctors and as well as managing the hospital resources and infrastructure. When you’re looking at a patient population, you’re not going to get 100 percent accuracy. But when it comes to a root cause, there are certain specific factors that helps predict the patient duration at the hospital.**

Some of the Data points that can help predict admission risks:

* Patients who take 5 or more medications
* Patient age and gender
* Patients with specific conditions including heart failure, sepsis or congestive obstructive pulmonary disorder



* [**Patient Admission Analytics**](http://www.fiercehealthcare.com/it/paris-hospitals-use-predictive-modeling-to-control-admissions) is the use of data analytics and machine learning to predict patient admissions down to the hour and is...
  + Built on top of Databricks Platform
  + Uses a machine learning **decision tree classifier** implementation to analysis a Patient Admission dataset
* This demo...
  + demonstrates a simple patient admission analysis workflow. We use Patient dataset from the [Kaggle Repository](https://github.com/jiunjiunma/heritage-health-prize/blob/master/modeling_set1.csv).

**Dataset:**

wget [https://raw.githubusercontent.com/jiunjiunma/heritage-health-prize/master/modeling\_set1.csv -O /tmp/patient/modeling\_set1.csv](https://raw.githubusercontent.com/jiunjiunma/heritage-health-prize/master/modeling_set1.csv%20-O%20/tmp/patient/modeling_set1.csv)

* Define schema with 142 columns for data ingestion

Schema:

MemberID\_t YEAR\_t ClaimsTruncated DaysInHospital trainset age\_05 age\_15 age\_25 age\_35 age\_45 age\_55 age\_65 age\_75 age\_85 age\_MISS sexMALE sexFEMALE sexMISS no\_Claims no\_Providers no\_Vendors no\_PCPs no\_PlaceSvcs no\_Specialities no\_PrimaryConditionGroups no\_ProcedureGroups PayDelay\_max PayDelay\_min PayDelay\_ave PayDelay\_stdev LOS\_max LOS\_min LOS\_ave LOS\_stdev LOS\_TOT\_UNKNOWN LOS\_TOT\_SUPRESSED LOS\_TOT\_KNOWN dsfs\_max dsfs\_min dsfs\_range dsfs\_ave dsfs\_stdev CharlsonIndexI\_max CharlsonIndexI\_min CharlsonIndexI\_ave CharlsonIndexI\_range CharlsonIndexI\_stdev pcg1 pcg2 pcg3 pcg4 pcg5 pcg6 pcg7 pcg8 pcg9 pcg10 pcg11 pcg12 pcg13 pcg14 pcg15 pcg16 pcg17 pcg18 pcg19 pcg20 pcg21 pcg22 pcg23 pcg24 pcg25 pcg26 pcg27 pcg28 pcg29 pcg30 pcg31 pcg32 pcg33 pcg34 pcg35 pcg36 pcg37 pcg38 pcg39 pcg40 pcg41 pcg42 pcg43 pcg44 pcg45 pcg46 sp1 sp2 sp3 sp4 sp5 sp6 sp7 sp8 sp9 sp10 sp11 sp12 sp13 pg1 pg2 pg3 pg4 pg5 pg6 pg7 pg8 pg9 pg10 pg11 pg12 pg13 pg14 pg15 pg16 pg17 pg18 ps1 ps2 ps3 ps4 ps5 ps6 ps7 ps8 ps9 drugCount\_max drugCount\_min drugCount\_ave drugcount\_months labCount\_max labCount\_min labCount\_ave labcount\_months labNull drugNull Readmitlabel

Readmit Label :

F.when(df1.DaysInHospital <= 0, 0.0).when(df1.DaysInHospital >= 1, 1.0).alias('Readmitlabel')

**Enrich the data to get additional insights to patient dataset**

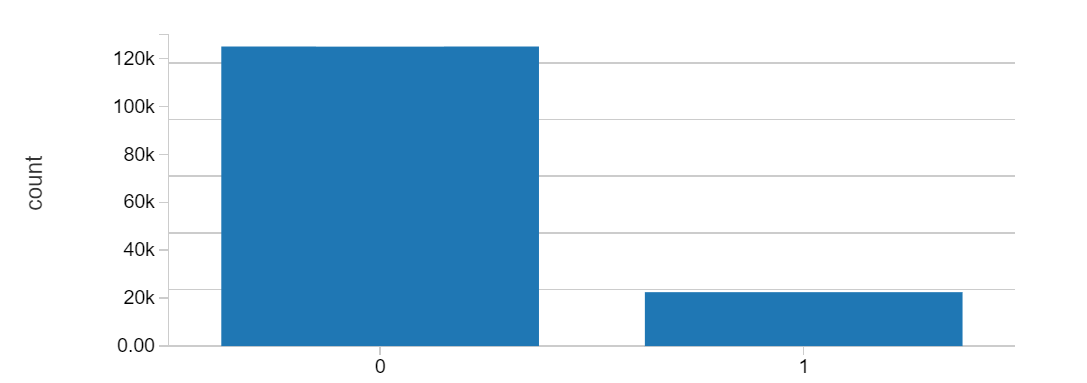
We count the number of data points and separate the churned from the unchurned

**Import SQL API Functions and Create table**

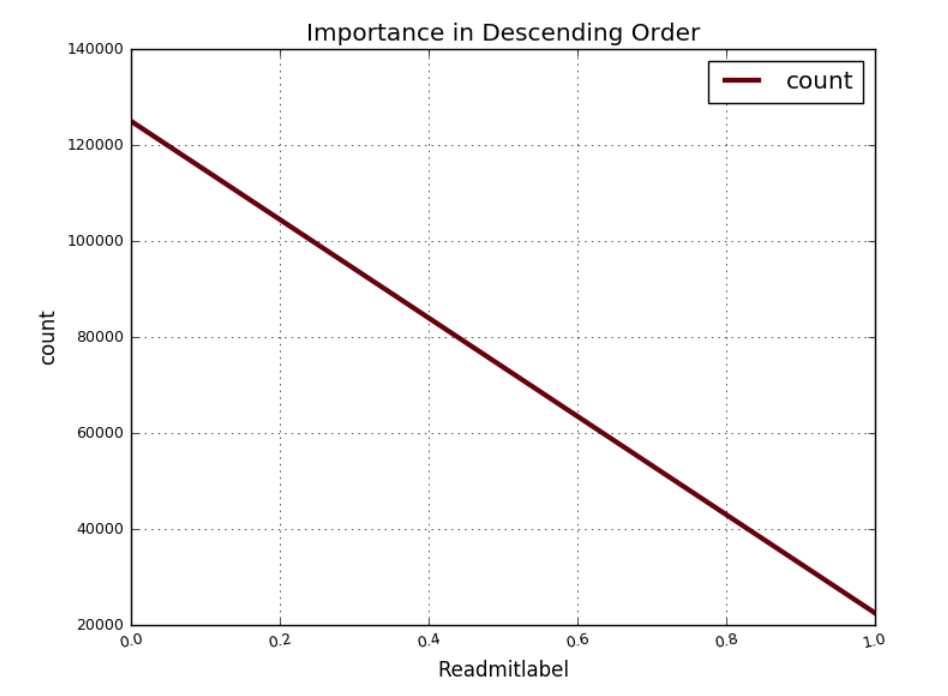
Age analysis of Patient DATA

### Explore Patient Data

### Breakdown of patient days in hospital using databricks graph

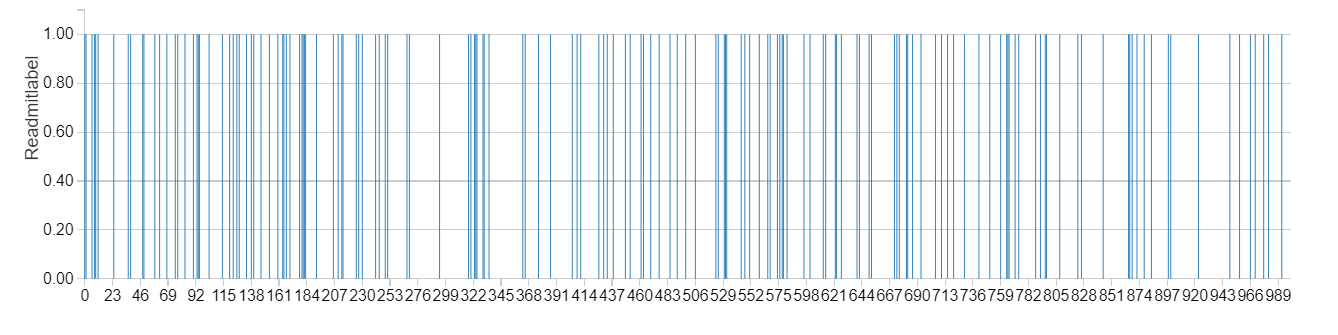


**Breakdown of patient days in hospital using matplotlib**



* Show the distribution of the account length.

display(df.filter(col('trainset') == 1).select("Readmitlabel"))



#### **Model creation by importing pyspark.ml.feature, import StringIndexer and VectorAssembler**

**And import ml pipeline,**

# We will use the new spark.ml pipeline API. If you have worked with scikit-learn this will be very familiar.

lrPipeline = Pipeline()

# Now we'll tell the pipeline to first create the feature vector, and then do the linear regression

lrPipeline.setStages([vecAssembler, aft])

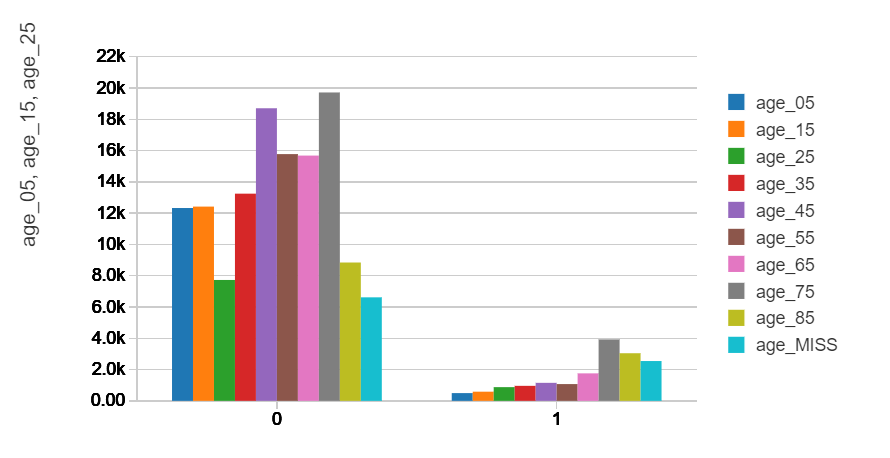
# Pipelines are themselves Estimators -- so to use them we call fit:

lrPipelineModel = lrPipeline.fit(finaldf)

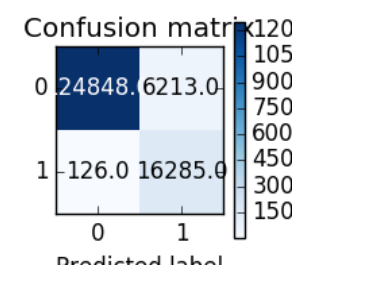
**Using Model for data predication,**

Defining the predictionsAndLabelsDF, predAnalysis and confusionMatrix.

**Visualizing the for-data prediction results,**



**Confusion matrix in matplotlib**



**Results Interpretation**

The plot above shows Index used to measure each of the patent admission days.



we have demonstrated prediction of number of days in hospital stay for a given patient based on decision tree classifier on Databricks Notebook.

We have achieved a very high accuracy of day in hospital prediction for a huge number of patients by just evaluating the general features like age, gender, types of drug and prescription used by the patient. This would help to better plan the infrastructure available to the patients in the hospital as well as the personnel management of doctors and nurses need as well.

At the end of the day, all this analysis and research is for one cause: To improve the quality of human lives. I hope this is, and will continue to be, the greatest motivation to overcome any challenge.