Springboard Capstone Project

DAN MATTHEWS

Health Care Costs in US

The cost of health care in the United States is a subject of concern, particularly with the changing landscape of the health insurance industry.

One particular point of interest is how health care costs vary depending on location.

- The cost of a given procedure can vary drastically depending on location.

For this report I studied the publicly available dataset 'Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) - FY2011‘ from data.gov

- Contains hospital charges for 3,337 health care providers across the US that receive IPPS payments for the top 100 most frequently billed Diagnosis Related Groups (DRGs) for FY2011.

Tasks:

- I characterize how these billed charges vary from hospital to hospital
- I attempt to predict whether or not the cost of a procedure (DRG) at a given hospital is above or below the national median cost for that procedure.

The Data

	DRG Definition	Provider Id	Provider Name	Provider Street Address	Provider City	Provider State	Provider Zip Code	Hospital Referral Region Description	Total Discharges	Average Covered Charges	Average Total Payments	Average Medicare Payments
0	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	DOTHAN	AL	36301	AL - Dothan	91	$32963.07	$5777.24	$4763.73
1	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	BOAZ	AL	35957	AL - Birmingham	14	$15131.85	$5787.57	$4976.71
2	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLORENCE	AL	35631	AL - Birmingham	24	$37560.37	$5434.95	$4453.79

Columns of note:

- **DRG Definition:** Code and description identifying the DRG. DRGs are a classification system that groups similar clinical conditions (diagnoses) and the procedures furnished by the hospital during the stay.
- **Average Covered Charges:** The provider's average charge for services covered by Medicare for all discharges in the DRG.
- **Average Total Payments:** The average of Medicare payments to the provider for the DRG, including the co-payment and deductible amounts that the patient is responsible for.
- **Average Medicare Payments:** The average payment to the provider just from Medicare.

Data Cleaning

To begin it was necessary to clean up the data to make it usable for the analysis.

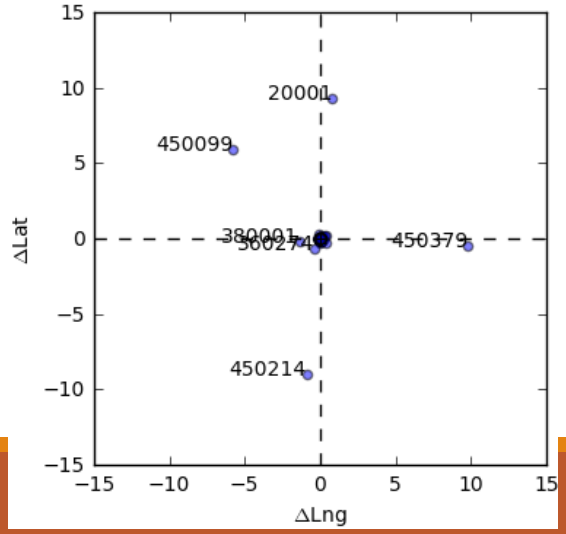
- Changed column name for easer coding (replace spaces with underscored, all lower case letters).
- Stripped dollar signs from monetary values and made them floats
- Padded zip codes to 5 character strings.
- Added a column (drg_id) that is a shorter string representing the DRG Definitions.
- Corrected truncated city names. In the csv file any provider who had a city name longer than 15 characters was truncated down to 15.

	drg_id	drg_definition	provider_id	provider_name	provider_street_address	provider_city	provider_state	provider_zip_code	hospital_referral_region_description	total_discharges	average_covered_charges	average_total_payments	average_medicare_payments
0	039	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	DOTHAN	AL	36301	AL - Dothan	91	32963.07	5777.24	4763.73
1	039	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	BOAZ	AL	35957	AL - Birmingham	14	15131.85	5787.57	4976.71
2	039	039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	10006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLORENCE	AL	35631	AL - Birmingham	24	37560.37	5434.95	4453.79

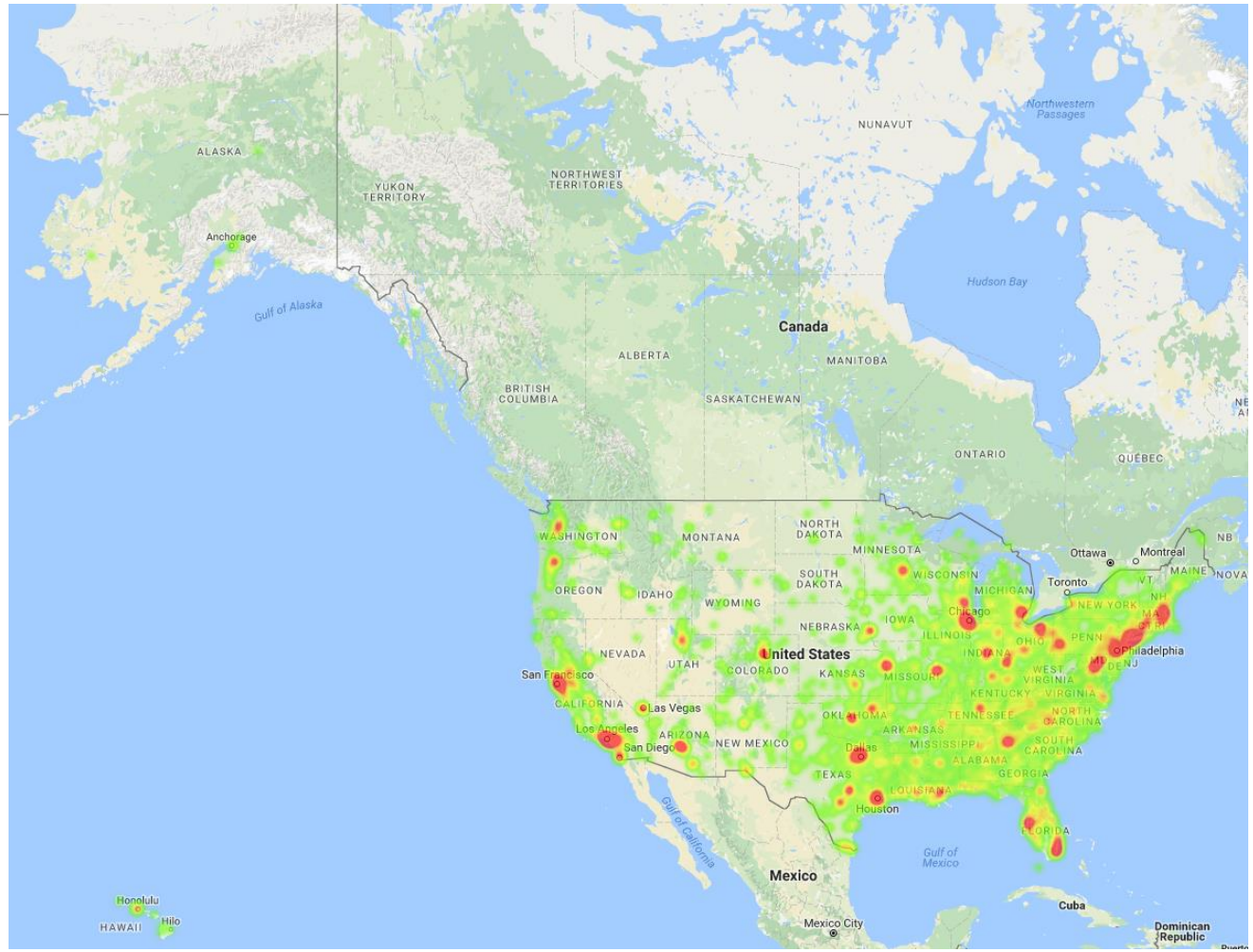
Geocoding

I used the address information provided in the dataset I used the GoogleMaps API to determine GPS locations of each provider.

- Combined all address columns into one string for querying the API.
- When providing the address information, sometimes giving the Provider Name as part of the address gave better results, other times NOT including it gave better results.
  - Ran both cases through the API and compared them to get the best results.



Heatmap of the provider locations



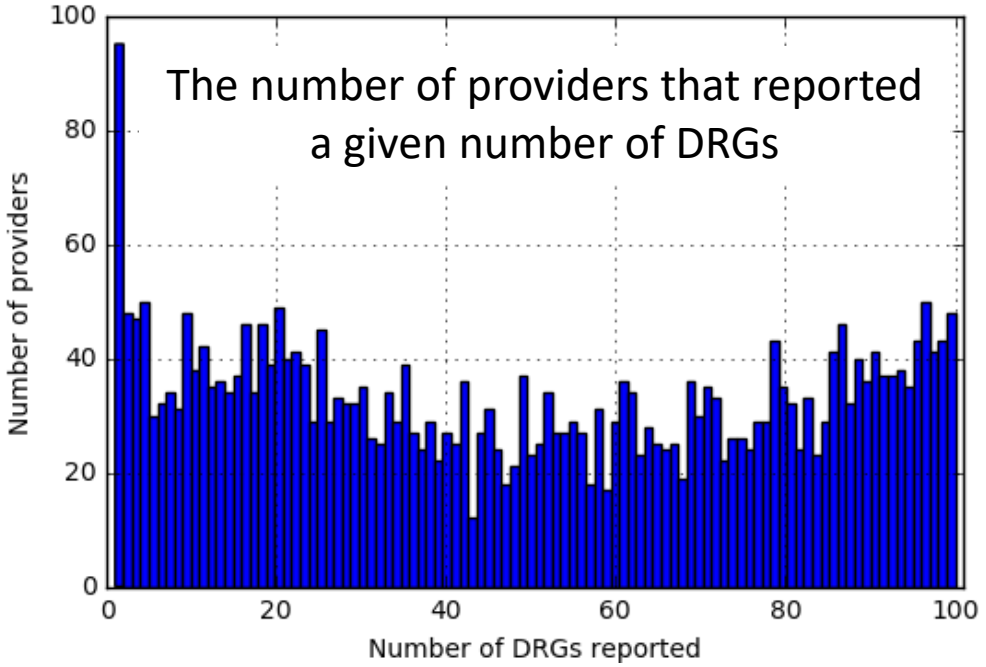
Exploratory Data Analysis

163,065 total rows

3,337 unique providers in all 50 states

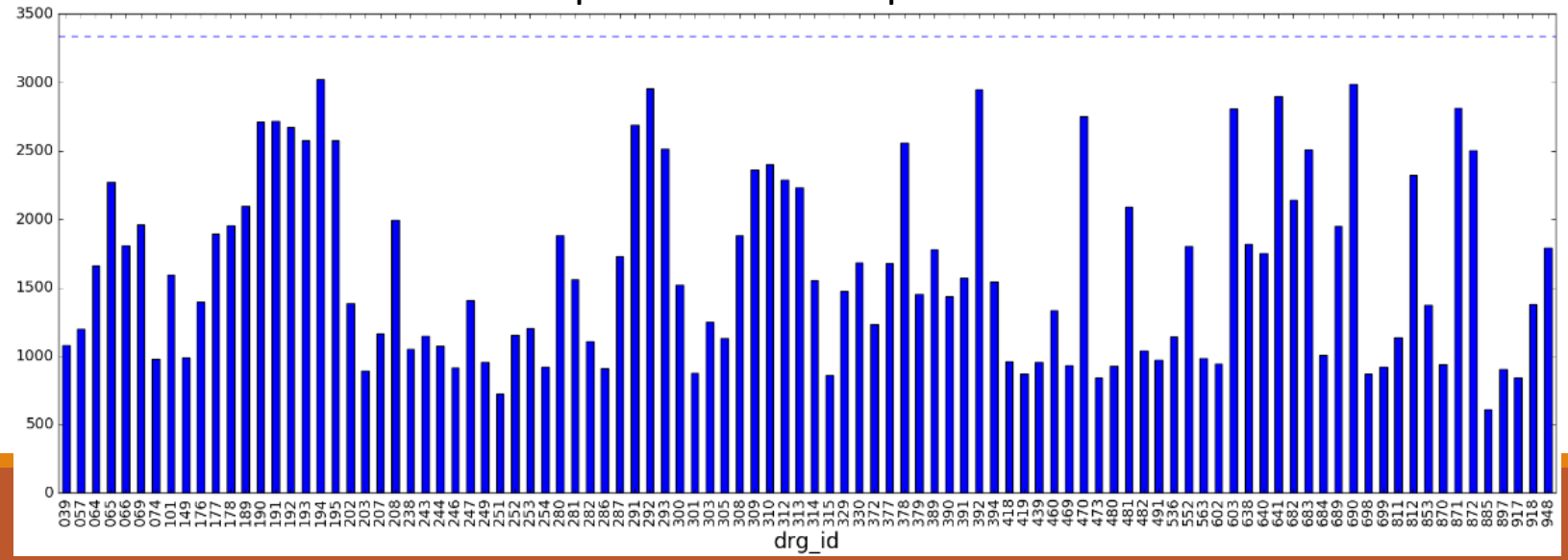
100 unique DRG Definitions

Most providers did not report data for all 100 DRGs



Example: 95 providers reported data for a single DRG, 48 reported data for all 100 DRGs

Number of providers that reported each DRG



Exploratory Data Analysis

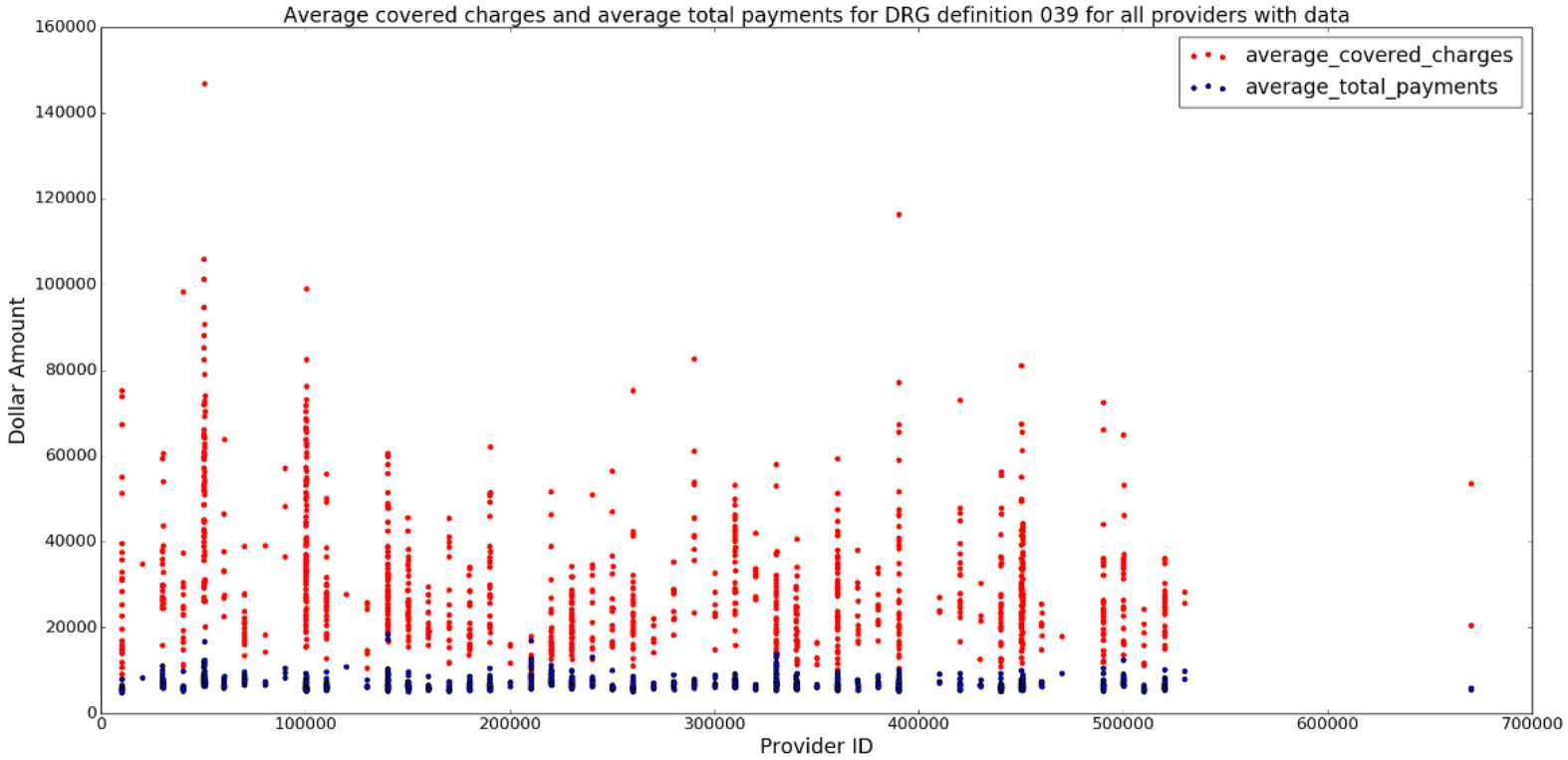
Average Covered Charges varied wildly depending on the provider.

For DRG 039:

- Provider ID vs dollar amount
- Red points: Average Covered Charges billed by the provider
- Blue points: Average Total Payments from Medicare to provider

Large variation in charges while payments are relatively consistent.

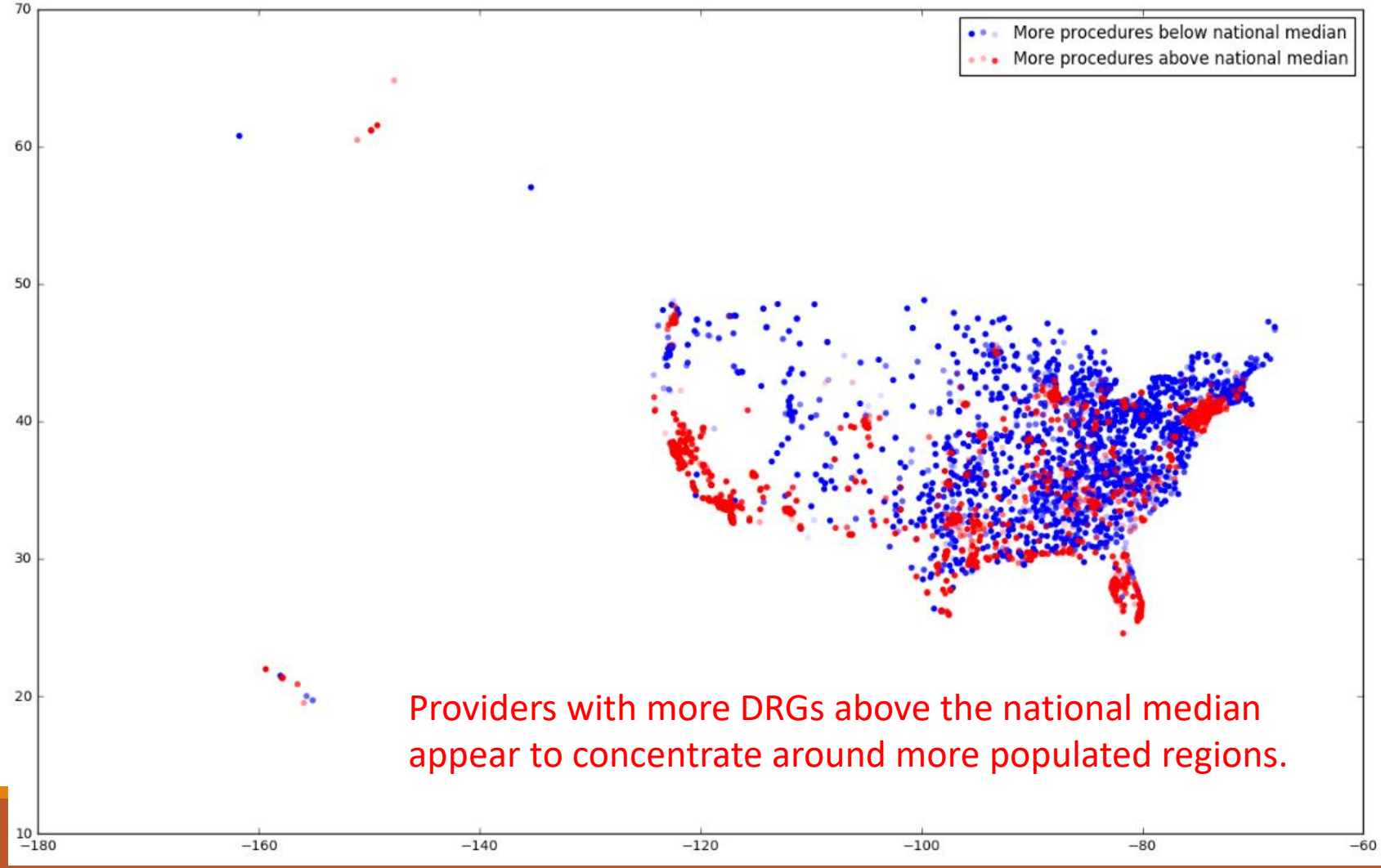
(Note: each vertical stripe is a different state due to the way they number the Provider IDs)



Exploratory Data Analysis

Looked at whether a provider tended to charge above or below the national median.

- Found national median charge for each DRG definition.
- Determined whether each provider had more DRGs above or below their national median.
- If they had more above, classified them as ‘Above’, otherwise classified them as ‘Below’.
- Combined this with the GPS coordinates of each provider to better visualize the data.



Providers with more DRGs above the national median appear to concentrate around more populated regions.

Machine Learning

Used the Random Forest Classifier from Scikit-Learn to predict whether or not the covered charges for a given DRG at a particular provider would be above or below the national median.

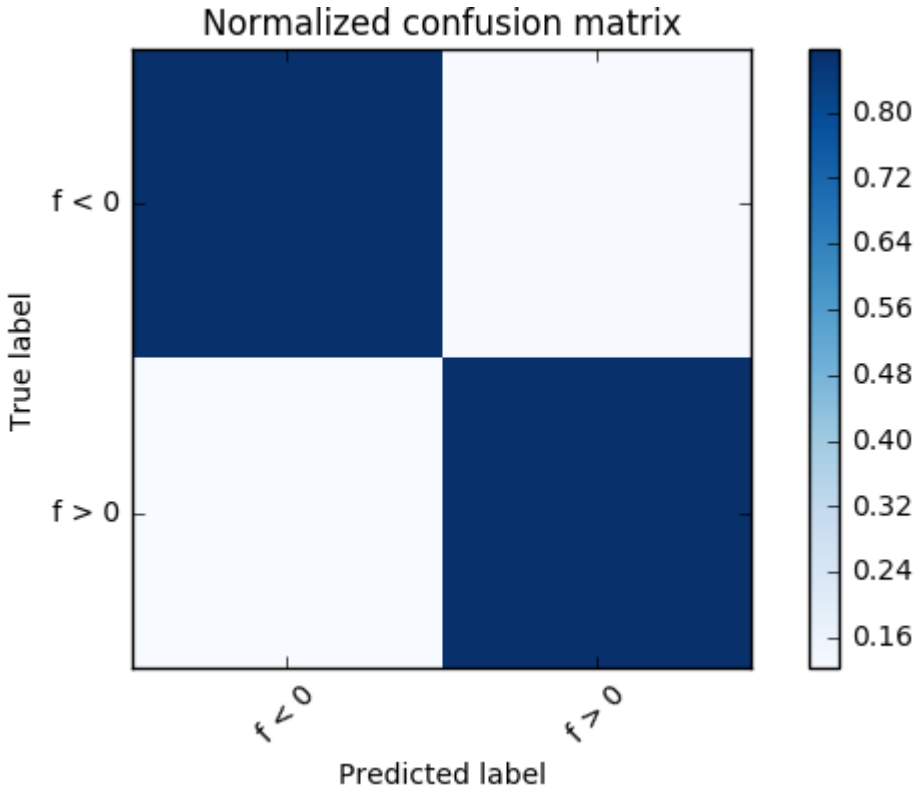
- For each row of the dataset I calculated the fractional difference of the Average Covered Charges as compared to the national median for that particular DRG.
  - $f \geq 0$: classified as 'True'
  - $f < 0$: classified as 'False'

- Features: the latitude and longitude of the provider
- Classification to predict: whether the covered charges billed by the provider are above or below the national median for that particular DRG Denfinition.

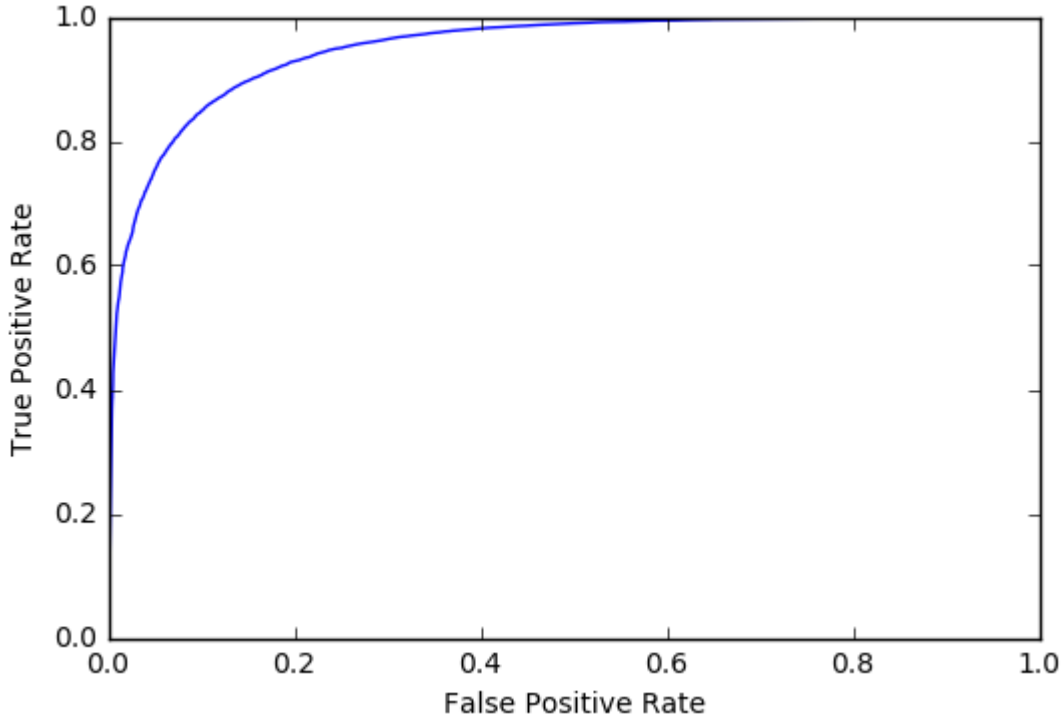
- Split the data into a training set (2/3 of data) and test set (1/3 of data)
- Trained the model on the training set
- Used the model to obtain predictions for the test set and compared those predictions to the actual results

Evaluation

Accuracy score: 0.87-0.88



AUC Score: 0.95



Conclusion

Prediction model worked well for predicting whether the cost of a procedure is above or below the national median.

Adding in other variables such as the population and median income by zip code seemed to have significant effect.

The number of hospitals within a given radius (e.g. 40 miles) as a feature seemed to have no effect either

For future analysis I would try to account for cost of living in the provider's region to see whether that would reduce the large variation in billed charges somewhat.