Capstone Report

Daniel Matthews

1 Introduction

The cost of health care in the United States is becoming a subject of more and more concern, particularly with the changing landscape of the health insurance industry. One aspect of this area of study is to look at the charges a hospital bills to the insurance company for a given procedure and how it can vary drastically from hospital to hospital. In this report I will be looking at the publicly available dataset 'Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) - FY2011' available from data.gov here: https://data.cms.gov/Medicare/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3. More details about this dataset can be found here.

To summarize, this dataset contains hospital charges for 3,337 health care providers in the United States that receive Medicare Inpatient Prospective Payment System (IPPS) payments for the top 100 most frequently billed Diagnosis Related Groups (DRGs) for Fiscal Year 2011. Using this dataset I characterize how these billed charges vary across the United States. In §2 I describe how I cleaned and made corrections to the data to make it ready for the analysis. In §3 I use the provider information to determine each provider's GPS location from the GoogleMaps API. In §4 I perform an Exploratory Data Analysis to determine some of the relevant information from the data such as the reporting statistics and how the cost varied by location. In §5 I use the Random Forest Classifier from the Python library Scikit-Learn to predict whether or not the cost of a procedure (DRG) at a given hospital will be above or below the national median.

2 Data Cleaning

This dataset provided some unique challenges to deal with in the analysis. First it was necessary to clean up the data in order to make it usable for the analysis. The data is read from the csv file and put it into a dataframe in pandas, where the first three rows of the dataset are shown below. I will describe what each column means throughout the analyses in this report.

	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

2.1 Preliminary Cleaning

There were a few changes I made to the dataset to make it easier to manipulate the data. For the each column name I replaced spaces with underscores and made all characters lowercase, which

makes writing code easier in the Jupyter notebook. For columns with dollar amounts I stripped the dollar sign from each string and converted them to float values so they can be manipulated mathematically. I also converted entries in the Provider Zip Code column from an integer to a 5 character string since they are never manipulated mathematically. This included padding 4 digit zip codes with a leading zero.

In addition to the changes described above I also added an additional column to the dataset. Since the DRG Definition column is a very long string it would be cumbersome to write out the long string every time I wanted to filter by procedure. To avoid this, I inserted a string type column at the beginning of the dataframe with just the three digit code describing the procedure (which is essentially the first three string characters of the DRG Definition).

2.2 City Name Correction

Since I will be attempting to gain GPS coordinate information for each provider from it's physical address, it is important for the address information of each provider to be correct. For some reason, in the csv file any provider who had a city name longer than 15 characters was truncated down to 15. For example, Havasu Regional Medical Center is located in Lake Havasu City, AZ, but in the Provider City column it's entry is 'LAKE HAVASU CIT'. In order to obtain to most accurate GPS coordinates over all providers it was necessary to correct this. These are the steps I took to make this correction:

1. I selected only the rows in the dataframe with unique city names, filtered on all cities with 15 characters and wrote these city names to a csv file.
2. I then edited this new csv file by hand.
   (a) Step 1 generated a list of 70 city names, which is feasible to edit by hand. After removing the cities whose names were exactly 15 characters and therefore not truncated, 46 city names remained.
   (b) For the 46 truncated names, I added a second column to the csv file that contains the corresponding full city name. So column 1 is the truncated city name and column 2 is the corrected city name.
3. For each row of this new csv file I iterate through the entire dataset and replace every occurrence of the truncated city name with the full city name.

After making these corrections I wrote the cleaned and corrected data to a new csv file that can be read in and used for future analysis. The first three lines of this new dataframe are shown below. The data dictionary in Appendix A gives a description of what each column means in more detail.

	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

3 Geocoding

In order to perform the analysis using the provider location in the United States it is necessary to find the GPS locations of all 3337 providers in the IPPS dataset. I used the GoogleMaps API to

accomplish this task.

3.1 Preparing the data for querying the GoogleMaps API

In order to query the GoogleMaps API the address must be given as a single string containing the full address. Since the dataset has the address information split up over multiple columns it was necessary to combine these columns into one single string column containing the full address. I read the Provider ID and address information (i.e. Provider Name, Provider Street Address, Provider City, Provider State and Provider Zip Code) into a dataframe and dropped all duplicate Provider IDs. Dropping the duplicates was necessary because in the full dataset there are multiple procedure types for a given provider. From this I created a new dataframe where I combined the address information columns into one string column for geocoding. I found that some characters created problems when querying the GoogleMaps API so I removed all commas and apostrophes from the address strings. Some of the provider names also ended with ", THE", so I removed those as well. I also added 'USA' onto the end of each address string to avoid getting GPS coordinates outside of the United States.

In the course of my analysis when querying the GoogleMaps API, I found that in some cases including the Provider Name field in the input address improved the geocoding results and in other cases including the name gave worse results. I first discovered this when I compared large outliers to what I found when looking up the address on the GoogleMaps website. To correct for this, I ended up querying the API using both cases (with and without the Provider Name) and compared them to improve the overall results, the details of which I describe later.The first two lines of each dataframe containing the address information are shown below.

	provider_id	address
0	10001	SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN AL 36301 USA
1	10005	MARSHALL MEDICAL CENTER SOUTH 2505 U S HIGHWAY 431 NORTH BOAZ AL 35957 USA

	provider_id	address
0	10001	1108 ROSS CLARK CIRCLE DOTHAN AL 36301 USA
1	10005	2505 U S HIGHWAY 431 NORTH BOAZ AL 35957 USA

3.2 Querying the GoogleMaps API

To query the GoogleMaps API, I defined a Python function that takes an input dataframe containing the Provider ID and Address, and finds the GPS coordinates for each address from the Google Maps API. The function outputs a dataframe containing the Provider ID, latitude and longitude of the provider. If the API returned a null result the latitude and longitude for that provider was designated with (lat, lng) = (0.0, 0.0). I inputted each dataframe described above (one where the address includes the Provider Name, one without) into this function and got back GPS locations for each provider. I then wrote each of these results into its own csv file.

To see the differences between the two cases I created a plot of the difference in longitude vs the difference in latitude for each provider. The plot is shown in Figure 1. The providers where the difference in either coordinate is larger than 0.5 degrees is labeled by provider ID. We can see that the distance is very large in some cases. To give an idea of the scale, 10 degrees of latitude is almost 700 miles!

To determine the best geocode results (i.e. GPS coordinates) for each provider I first listed out all of the providers where the differences in the two results are above some threshold in degrees. I

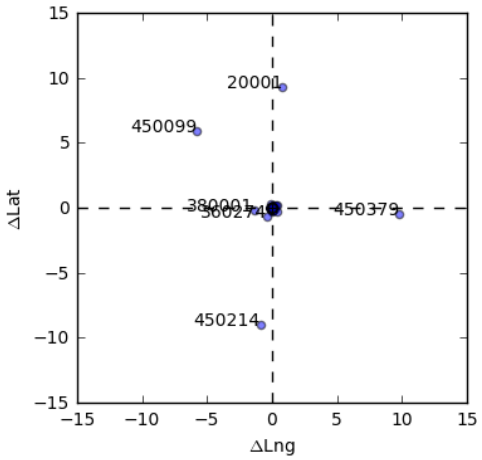


Figure 1: The difference in latitude and longitude (in degrees) between the two cases of including and not including the Provider Name in the address. The points with large differences are labeled by their Provider ID.

then searched using the address of those providers on the GoogleMaps website and used the GPS coordinates given there to determine which result is the correct one. Doing this by hand was a bit tedious, but it was not an overwhelming number of cases so it was the quickest way to get it done. The threshold I chose was 0.05 degrees, which gave 67 different providers. In latitude 0.05 degrees is about 3.5 miles. I found that for the cases where the difference was very large, sometimes including the provider name gave the best results and sometimes not using the name was best. As the difference got smaller, NOT including the provider name consistently gave the better results. For this reason I was comfortable only checking down to the threshold of 0.05 degrees, and for all cases below that threshold I just kept the results from not using the provider name.

I compiled all of the best results for each provider into a single csv file which can be used for further analysis. After reading these coordinates into a dataframe I generated a heatmap showing every provider overlaid on a map of the United States. This is shown in Figure 2.

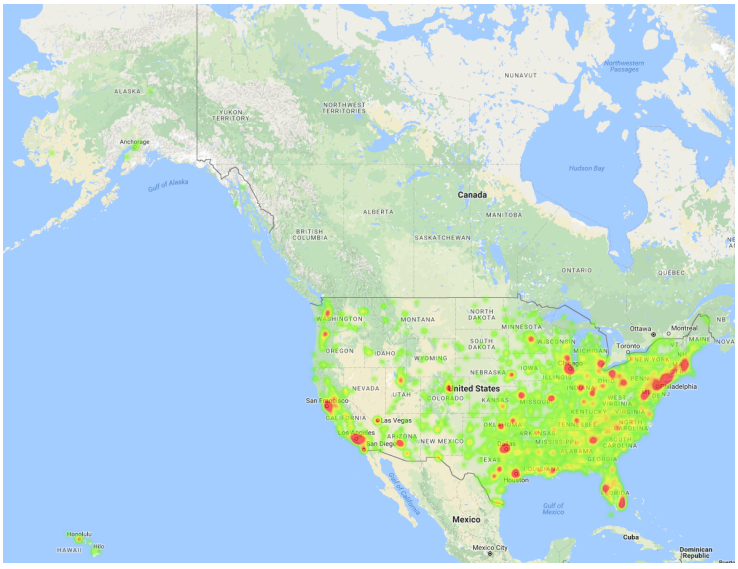
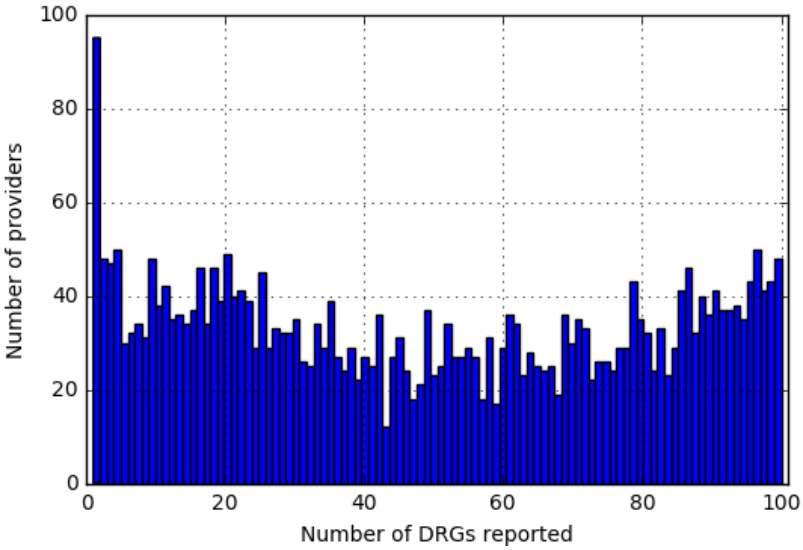


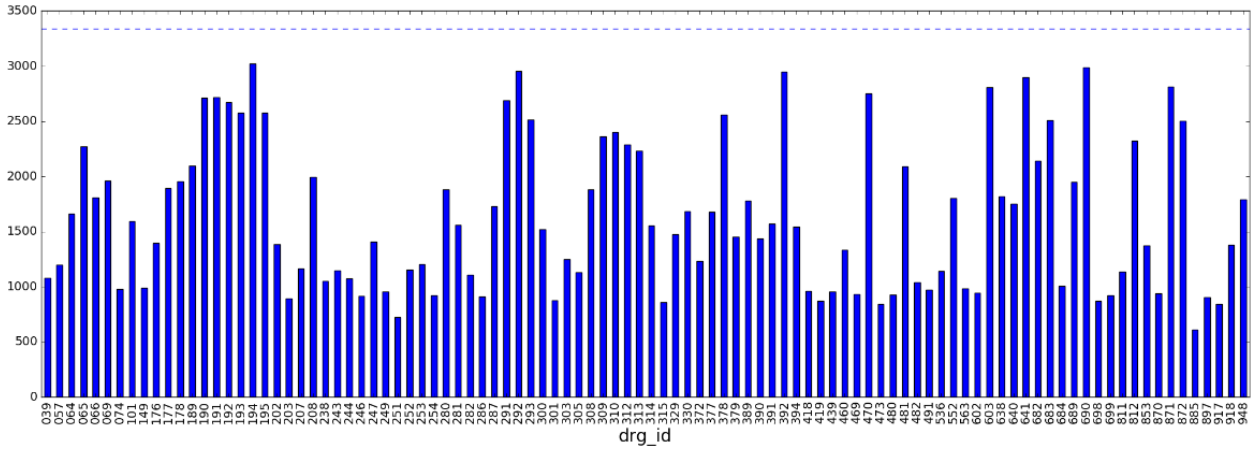
Figure 2: A heatmap of every provider in the IPPS dataset.

4 Exploratory Data Analysis

Next I performed some Exploratory Data Analysis (EDA) to get a general idea of the kinds of numbers in the dataset. In total there are 163,065 rows in this dataset. It contains data from 3337 providers in all 50 states. Each provider has a Provider ID number, provider name, address and hospital referral region. Each provider has reported billed cost and payment data for a number of procedure categories, i.e. DRG Definitions. There are 100 different DRG Definitions in this dataset across all providers, although most of the providers did not provide data for every single DRG. To get an idea of the number of DRG Definitions reported over all providers, Figure 3a shows the number of providers that reported a given number of DRG Definitions. For example, 95 providers only reported data for a single DRG definition, and 48 providers reported data for all 100 DRGs. Figure 3b shows the number of providers that reported each DRG Definition.



(a)



(b)

Figure 3: (a) The number of providers that reported a given number of DRG Definitions. (b) The total number of providers nationwide that reported data for each DRG definition. The blue dashed line is the total number of unique providers in the data set (3337).

For each given DRG Definition each provider reported:

- **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.

These averages are over the total discharges that fell under the particular DRG Definition.

To see how these values vary across providers I plotted the Average Covered Charges and Average Total Payments for a given DRG Definition as a function of the Provider ID. Figure 4 shows this plot for DRG Definition 039. We can see that the average charge billed by the provider to Medicare varies wildly from provider to provider, while the average payments made to the provider are much more consistent. (Note: The discrete nature of the plot in the x-direction (Provider ID) comes from how the providers are numbered by state. Essentially each of those stripes in the y-direction represent providers in a different state.) The other DRG Definitions showed similar trends.

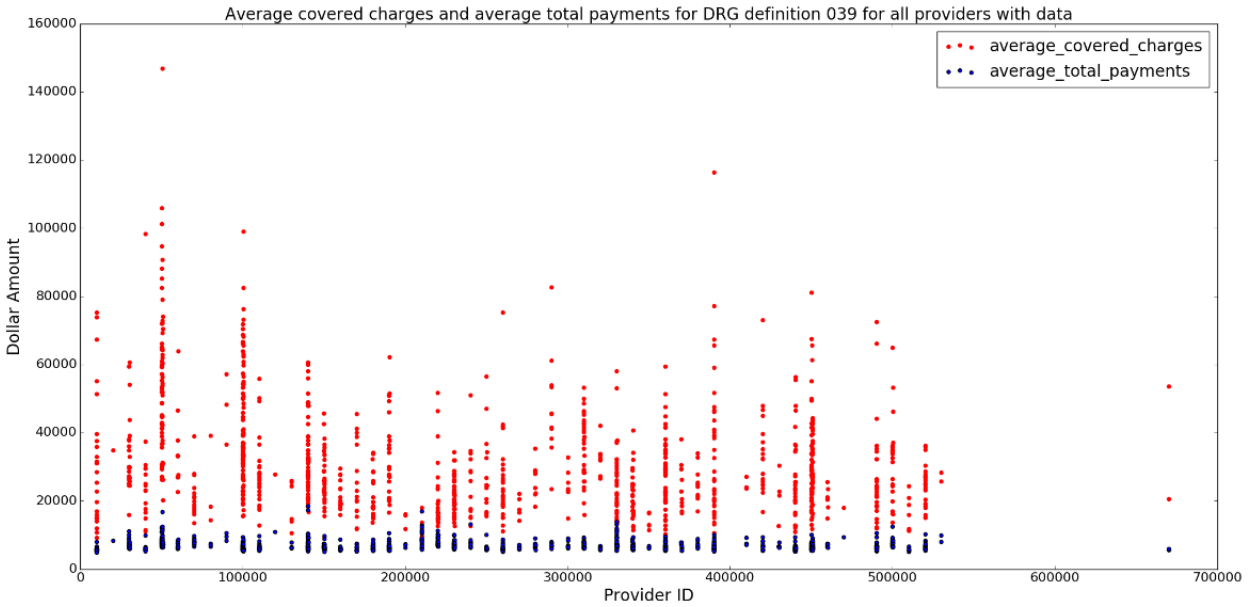


Figure 4: A scatter plot of the Average Covered Charges and Average Total Payments for DRG Definition 039 as a function of the Provider ID.

Another way to characterize these numbers is to look at their median values over all providers nationwide. Figure 5 shows the national median cost, median payments and median Medicare payments for each DRG definition. We can see that the median payments and Medicare payments are consistently a fraction of the covered charges as was shown in the previous plot.

Next I looked at the fractional difference in total charges, payments and Medicare payments from the national median for each row in the dataset. So for each DRG Definition at each provider, I subtracted the monetary values from their corresponding national median and divided by the national median. This gives a quantity that describes how much the provider charges or receives as compared to the national median for that DRG Definition. Figure 6 shows a histogram of these fractional differences.

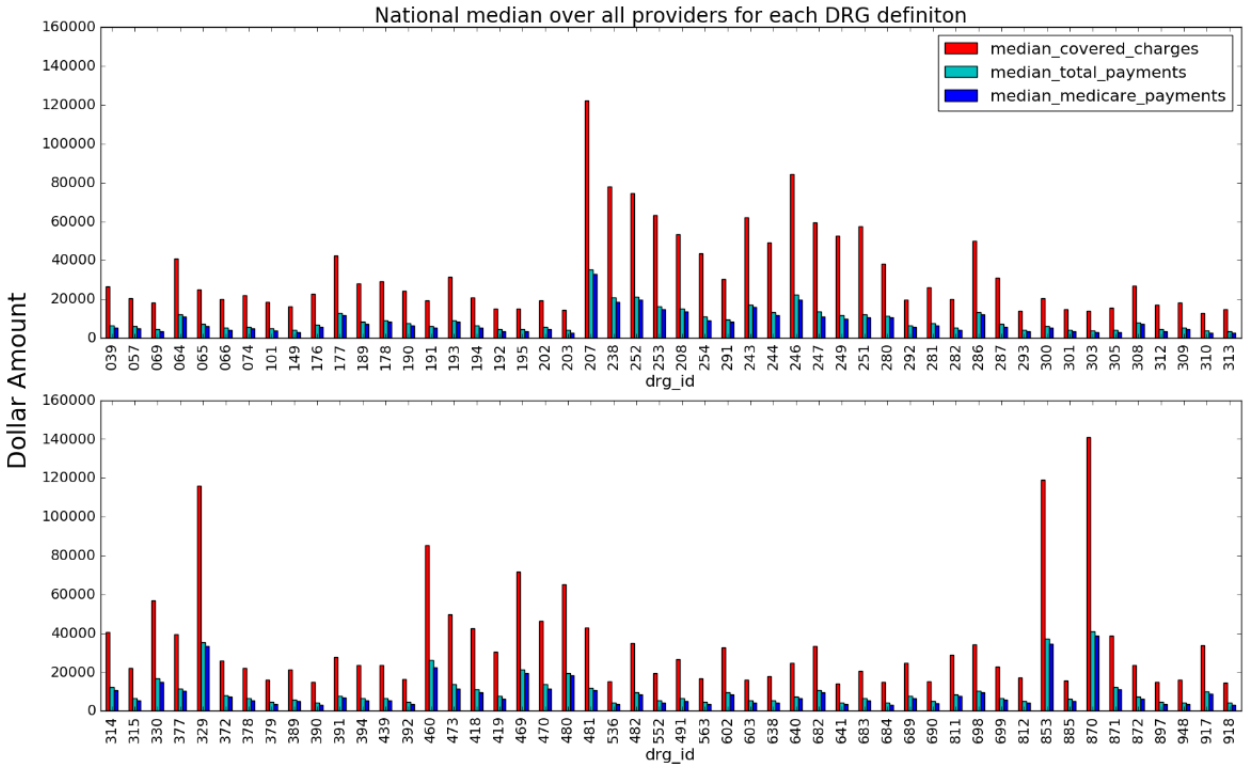


Figure 5: The national median values of the Average Covered Charges, Average Total Payments and Average Medicare Payments for each DRG Definition,

Since the Average Covered Charges seems to vary drastically depending on the provider I wanted to look at whether a provider tended to charge above or below the national median for a procedure. To characterize this, for each provider I determined whether they had more DRG Definitions above or below the national median. Since most providers had reported charges for multiple DRG Definitions I simply added up how many of them were above and how many were below the national median. If the provider had more than half of its DRG Definitions above the national median I classified it as 'Above', and if more than half were below I classified it as 'Below'.

To better visualize this data I combined this classification with the GPS coordinates of each provider as determined in section 3. I then plotted the coordinates of each provider, where each provider was labeled as either 'Above' or 'Below'. Figure 7 shows the coordinates of each provider, where the providers classified as 'Above' are red and the providers classified as 'Below' are blue. From this plot it appears that the providers who tend to charge above the national median are in areas of high population density such as in large cities. This could be expected due to the difference in cost of living in these areas.

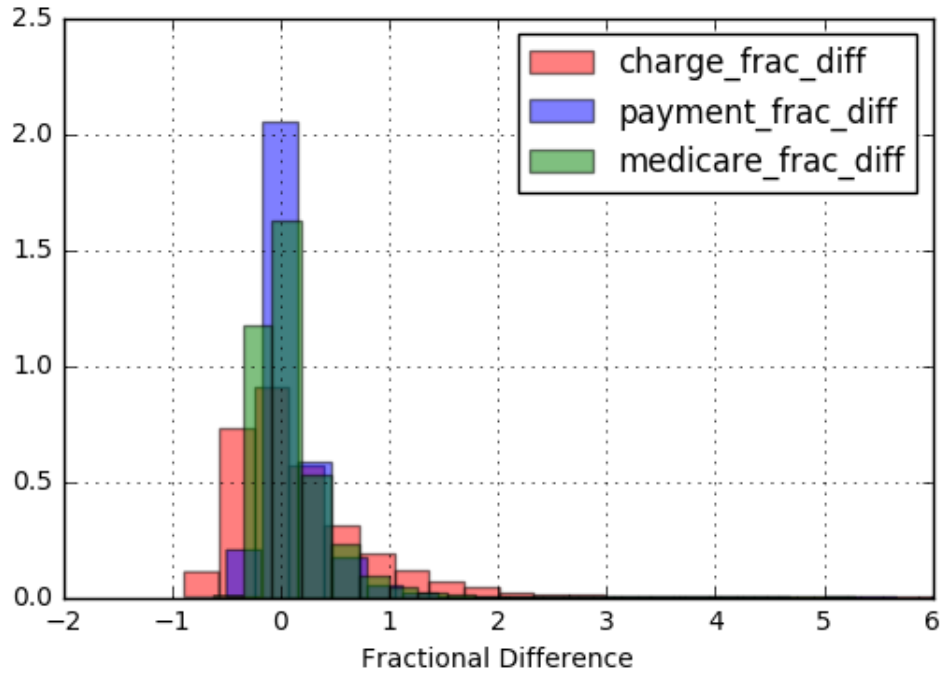


Figure 6: Histogram of the fractional differences from the national median for the Average Covered Charges, Average Total Payments and Average Medicare Payments.

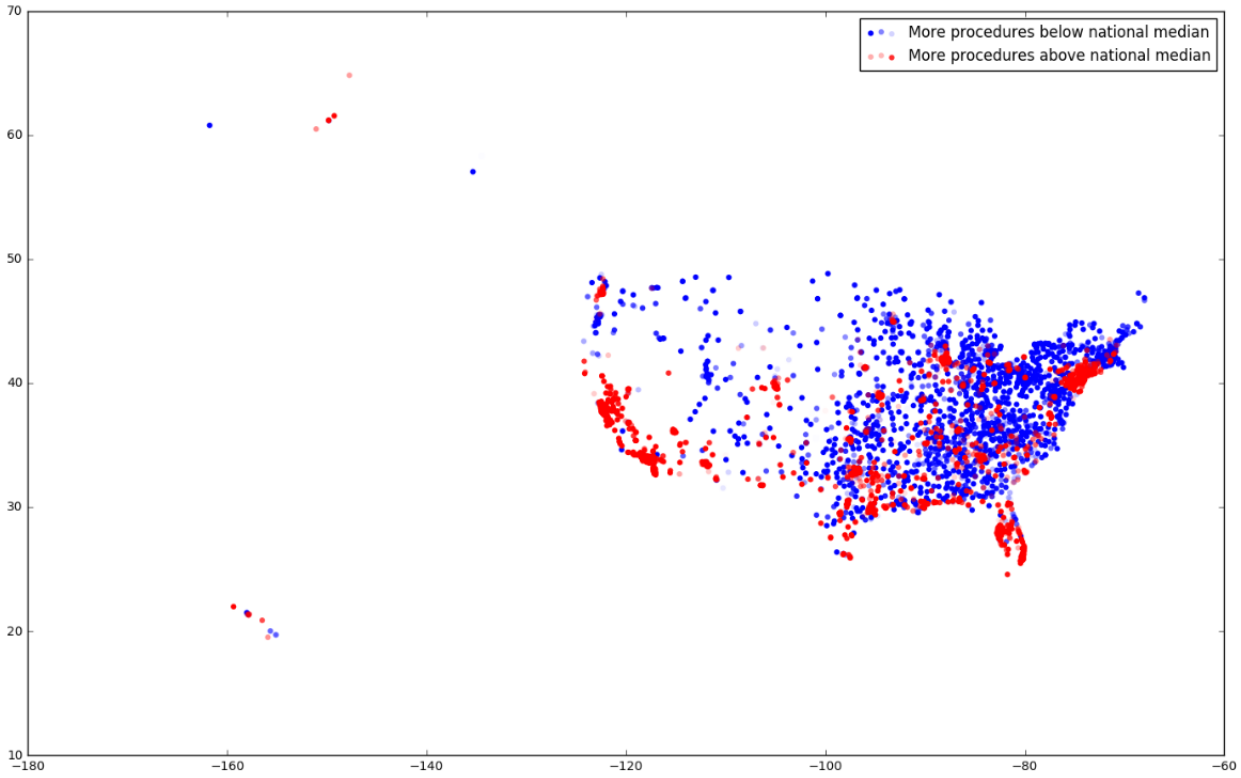


Figure 7: A scatter plot of all providers in the data set, where providers that tend to charge above the national median are red and providers that tend to charge below the national median are blue.

5 Machine Learning: Random Forest Classifier

Since the Average Covered Charges varied so much depending on the provider I thought it would be interesting to see if I could use the dataset to predict these billed charges for a procedure that falls under a DRG Definition at a given provider. It turned out to be quite difficult to predict the actual monetary value with very much accuracy, so I decided to see if I could predict whether or not the covered charges would be above or below the national median for that particular DRG. This made the classification of the cost more coarse but it did give better results.

5.1 Prediction Model

For each line of the dataset I calculated the fractional difference of the Average Covered Charges compared to the national median for that particular DRG Definition. If this fractional difference was greater than or equal to zero (above or equal to national median), I classified that DRG at that particular provider as 'True'. If the fractional difference was less than zero (below the national median), I classified it as 'False'. The features I used to make the prediction were the GPS coordinates of each provider. I thought that location might be an indicator given what I discovered in Figure 7 in Section 4. So for each classification ('True' or 'False') of a DRG at a given provider I assigned the GPS coordinate of that provider. So in summary:

- **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 Definition.

To attempt to predict the classification of a DRG at a given provider I used the Random Forest Classifier from the ensemble method in the Scikit-learn Python library. I initially ran the RFC on the entire dataset to see how well I could predict the classification and did get an accuracy score of 0.88, but this does not actually demonstrate whether or not the model is a good predictor. This is because you can have issues such as overfitting the data when using the full dataset. To see how robust the model is and whether or not it generalizes well I split the data into a test set and training set and used the training set to train the model and then tested the model by making predictions on the test set. The training set was constructed by randomly selecting 2/3 of the data, with the remaining data acting as the test set. The predictions are compared to the actual data to see how well the model performs. When I train on the training set and apply the model to the test set, I obtain an accuracy score of $\sim 0.87 - 0.88$, which is a very similar score compared to when I used the full dataset.

5.2 Evaluation

To evaluate the quality of the model predictions I looked at the confusion matrix, which is a visualization of the rate of correct predictions (true-positives and true-negatives) and incorrect predictions (false-positives and false-negatives). Figure 8 shows a normalized confusion matrix, which gives the rates as a fraction. The first row shows how many correct and incorrect predictions for classifications of charges below the national median, and the second row shows the prediction rates for classifications of charges above the national median. The confusion matrix is symmetric, so in both cases the rate of correct predictions are around $\sim 0.87 - 0.88$.

I also looked at the Receiver Operating Characteristic (ROC) curve for this model. The ROC curve plots the false positive rate vs the true positive rate as the probability threshold for discriminating between the two classes is varied. Ideally you want a low false positive rate and high true

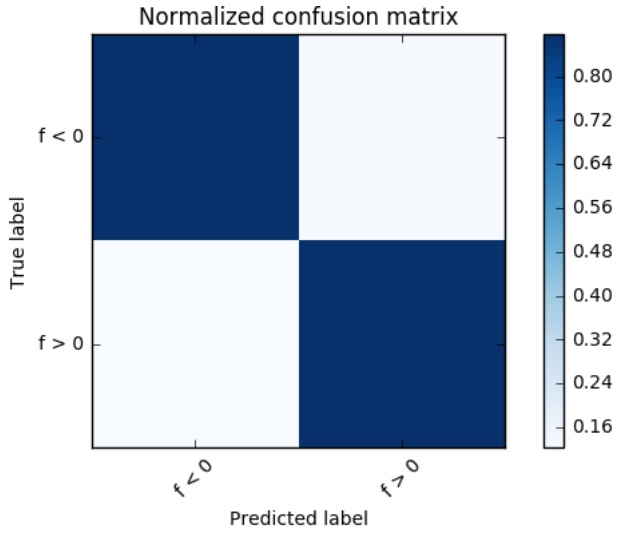


Figure 8: Confusion matrix.

positive rate, so the upper left corner of an ROC plot is preferred. Figure 9 shows the ROC curve for the predictor model. For the varying probability thresholds the curve is in the upper left corner, which is a good result. To evaluate how well the model does from the ROC curve I calcluated the Area Under the Curve (AUC) score. The AUC score for this curve is 0.95, which shows it is an effective predictor.

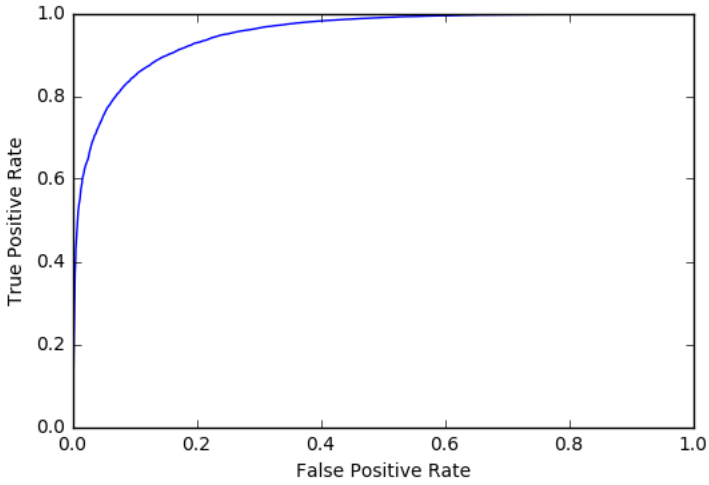


Figure 9: ROC curve.

5.3 Additional Features

Adding in other variables such as the population and median income by zip code seemed to have no significant effect. I also calculated the number of other providers within 40 miles of each provider, but adding that variable had no effect as well. The number of providers within a certain radius is likely related to the population density. I also found that I get the same results with just the Provider ID tag as the input X using pandas get_dummies to convert the variable to an indicator variable rather than numerical value. This makes me think that the GPS coordinates are essentially

acting as an indicator for the specific provider rather than giving any useful information as far as relative position to other hospitals.

A Data Dictionary

Name	Type	Description
drg_definition	String	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.
provider_id	Integer	Provider Identifier billing for inpatient hospital services.
provider_name	String	Name of the provider.
provider_street_address	String	Street address in which the provider is physically located.
provider_city	String	City in which the provider is physically located.
provider_state	String	State in which the provider is physically located.
provider_zip_code	String	Zip code in which the provider is physically located.
provider_hospital_referral _region_description	String	Hospital referral region in which the provider is physically located.
total_discharges	Integer	The number of discharges billed by the provider for inpatient hospital services.
average_covered_charges	Float	The provider's average charge for services covered by Medicare for all discharges in the DRG. These will vary from hospital to hospital because of differences in hospital charge structures.
average_total_payments	Float	The average of Medicare payments to the provider for the DRG, including the DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included are co-payment and deductible amounts that the patient is responsible for.
average_medicare_payments	Float	The average payment to the provider just from Medicare.