Prescription Based Prediction

Aviv Yaish, Chaim Hoch

Introduction

The Centers for Medicare & Medicaid Services (CMS) is a federal agency within the United States Department of Health and Human Services (HHS) that administers the Medicare program and works in partnership with state governments to administer Medicaid, the State Children's Health Insurance Program (SCHIP), and health insurance portability standards.

Medicare Part D, also called the Medicare prescription drug benefit, is a United States federal-government program to subsidize the costs of prescription drugs and prescription drug insurance premiums for Medicare beneficiaries. In 2015, CMS publicly released a dataset of prescriptions made under Medicare Part D in 2013.

This dataset includes the number of drugs given by each provider and anonmized information about each provider. This dataset can be used in order to gain meaningful information about prescription habits of providers, as well as statistical information about the medical providers.

General Statistics

The dataset is a pairing between a list of prescribed drugs and anonymized data about the provider, and over all includes 239,930 records. The list of drugs includes drugs prescribed more than 10 times during the year the data was collected, and the number of prescriptions per drug. These drugs include brand names and generic names alike.

For each provider, the dataset includes information such as the gender of the provider, specialty, region, years practicing, location type and the providers' National Provider Index (NPI).


```
{"cms_prescription_counts":
    {"CEPHALEXIN": 28,
     "AMOXICILLIN": 73,
     "CLINDAMYCIN HCL": 11},
 "provider_variables":
    {"settlement_type": "non-urban",
     "generic_rx_count": 112,
     "specialty": "General Practice",
     "years_practicing": 7,
     "gender": "M",
     "region": "Midwest",
     "brand_name_rx_count": 0},
 "npi": "1578587630"}
```


Figure 1: Data point example

After looking at the 'specialty' field, it was observed that it has a 'long tail' (see Figure 2), meaning there are many specialties which have a small quantity of providers who practice that specialty. When trying to gain meaningful insights about the specialty, this might cause problems which we wanted to avoid.

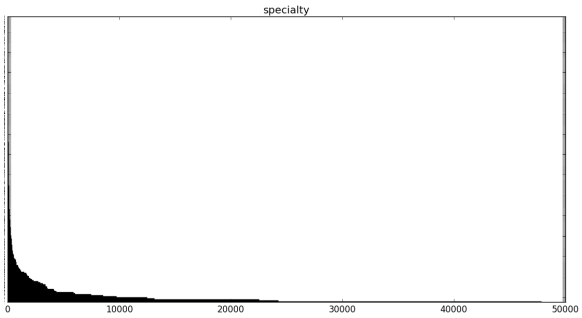


Figure 2: Specialty long tail

In order to solve this problem, we eliminated providers who specialized in fields that have less

than 50 providers, as well as doctors who prescribed less than 30 prescriptions. This left 55357 entries and got rid of the long tail (see Figure 3), but kept the same statistical relations (see Figures 4, 5).

Diving deeper into the data, we can see that gender isn't distributed evenly according to all traits. For example, there are disproportionately more women among newer doctors, while there are much more men among veteran doctors (see Figure 6a). Also, there are specialties which are dominated by different genders (see Figure 6b).

Hopefully, these skewed distributions will be easy to detect and will help us to predict the various traits.

Goal and Methods

Thanks to the CMS, we have a plethora of data. Besides obtaining interesting information from the various visualizations of the data, it would be interesting to explore the various

Prediction of prescriber traits from prescriptions

Goal: to predict the various traits of each prescriber from the list of prescriptions he has given.

Uses: If we are able to predict with a good accuracy a certain trait, for example gender, it means that the distribution of that trait according to medication is not uniform. This allows us to find certain predispositions - maybe male doctors have a tendency to prescribe certain medications, or maybe inexperienced doctors prescribe more medications than experienced.

Thus, being able to predict with high accuracy a certain trait means that maybe people in the field should have a look, find the causes for that certain predisposition, and understand if it should me mended.

Methods: there are well known methods of prediction, and we've used the following:

1. Logistic Regression
2. Adaboosted Logistic Regression
3. Random Forest
4. Adaboosted Random Forest
5. Bagged Random Forest
6. SVM
7. Multi-Layer Perceptron

In the results section we will detail the accuracy achieved by each method.

The code for this is contained in the "classifications" method and it's various subroutines.

Modeling specialties according to prescriptions

Goal: to develop a method to model specialties from prescriptions.

Uses: Given such a model, it would be interesting to look at "oddities" - medications that were developed for certain specialties, but that according to the model should also be grouped under a different specialty. These might be off label medications used by certain perscribers, and if these medications are not known as such by the medical community, might require further research as to their efficacy when used off label.

Method: We treated each list of prescriptions by a single prescriber as document, and each specially as a topic. Then, by performing LDA on all the prescriptions (in essence, our "corpus"), we get an assignment of specialty to each medication. Going through these topics, we look at each specialty and look for the one that maximizes the document topic distribution of the corpus as produced by the LDA. This maximizing specialty is added to a counter. Eventually, the counter produces a sorted list of the most likely specialties for each topic.

The code for this is contained in the "learn_specialties" method and it's various subroutines.

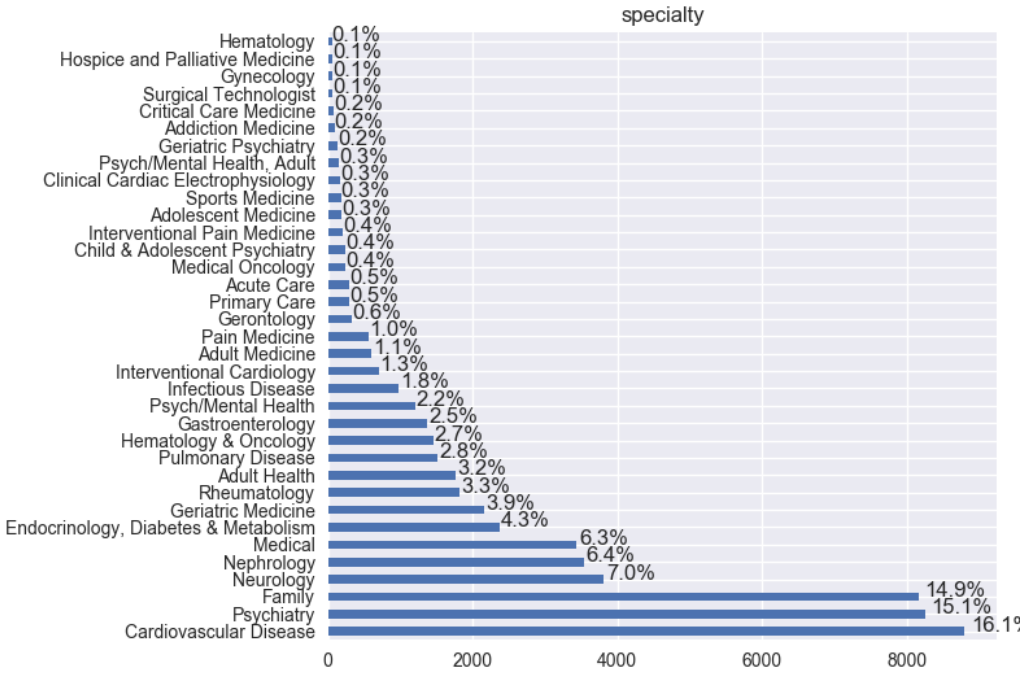


Figure 3: Specialty long tail, after removal

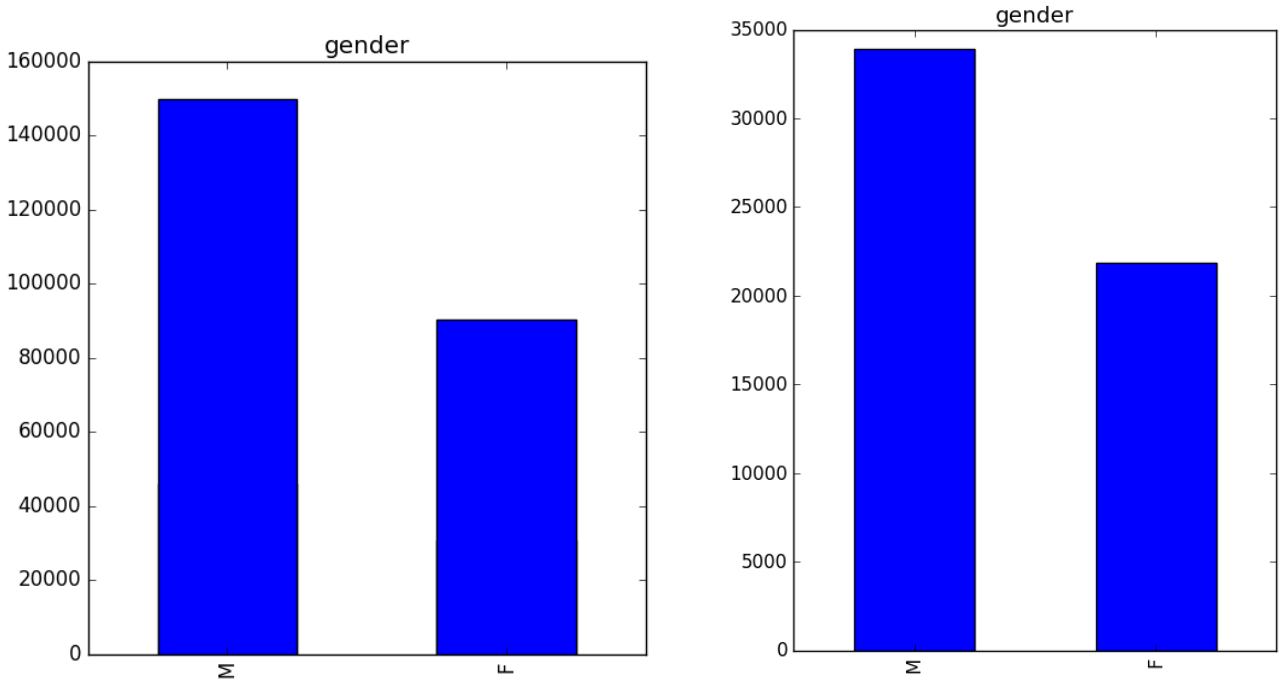


Figure 4: Gender split, before (left) and after (right) removal

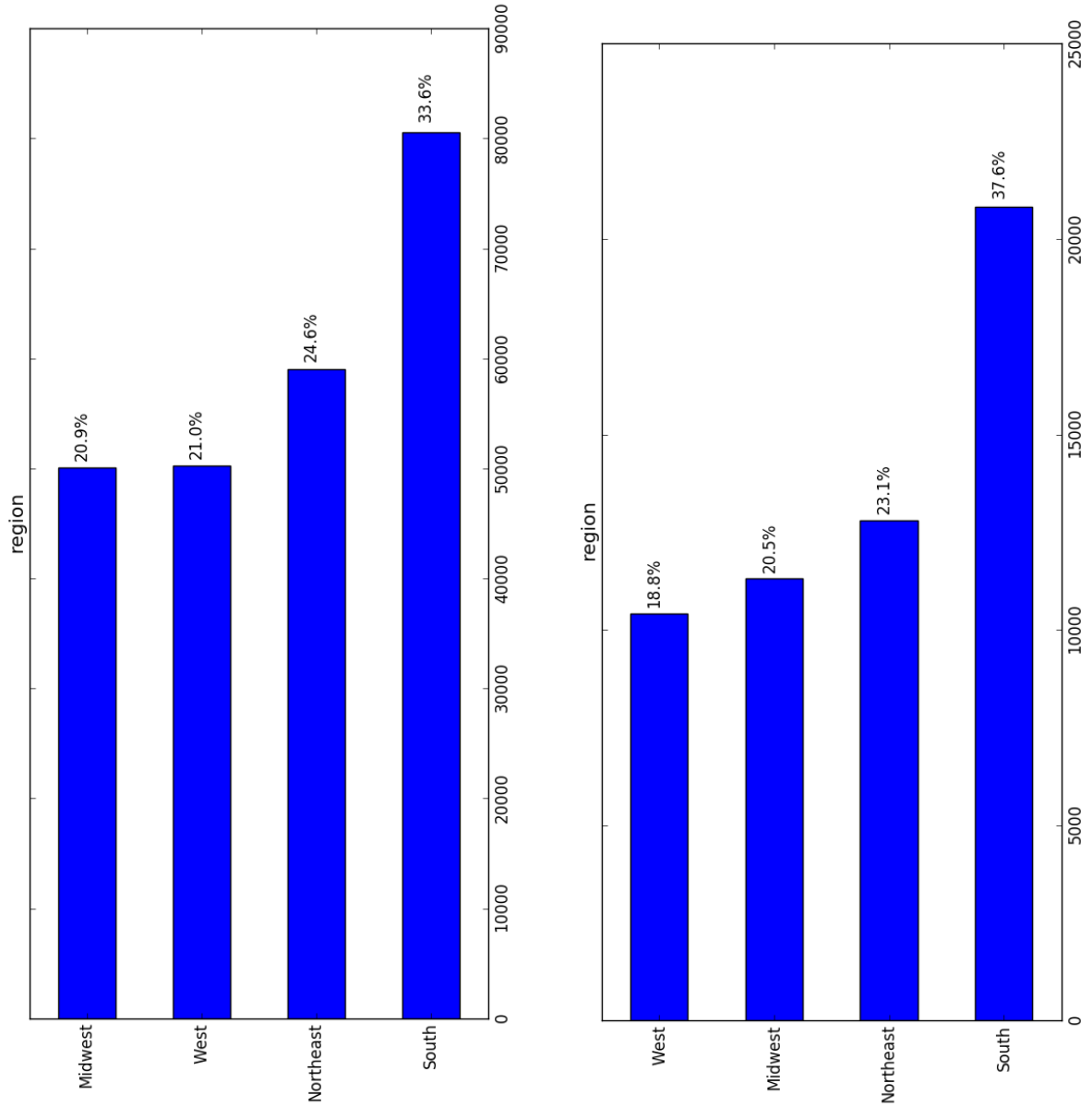
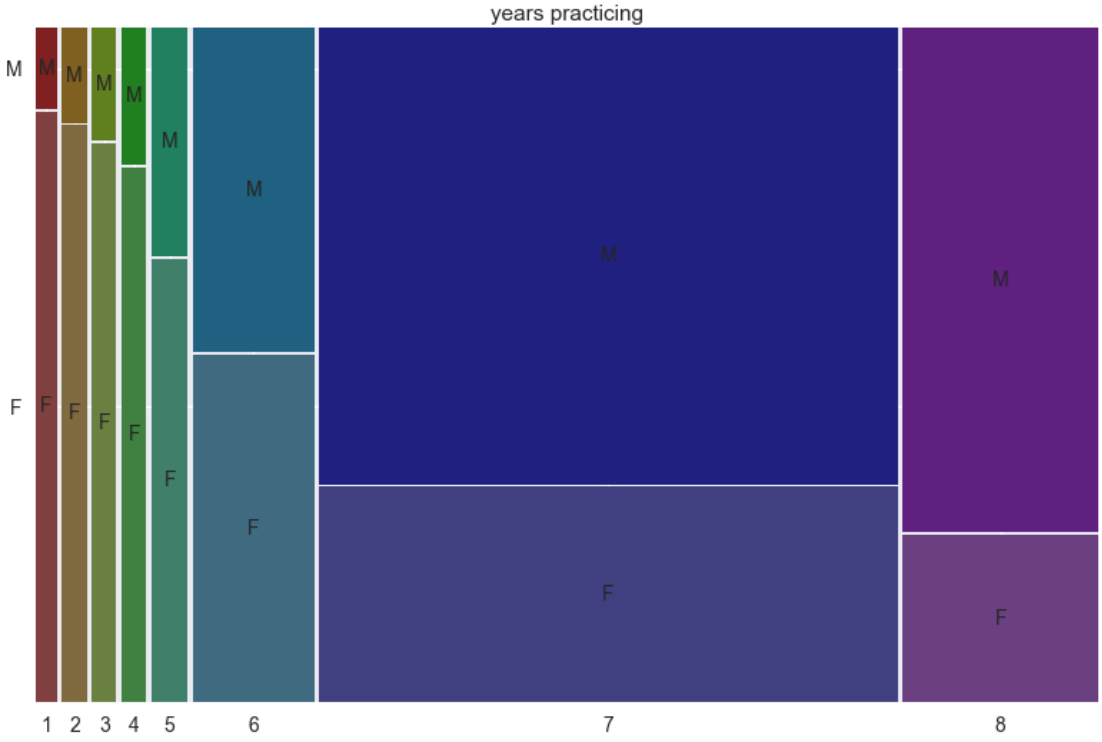
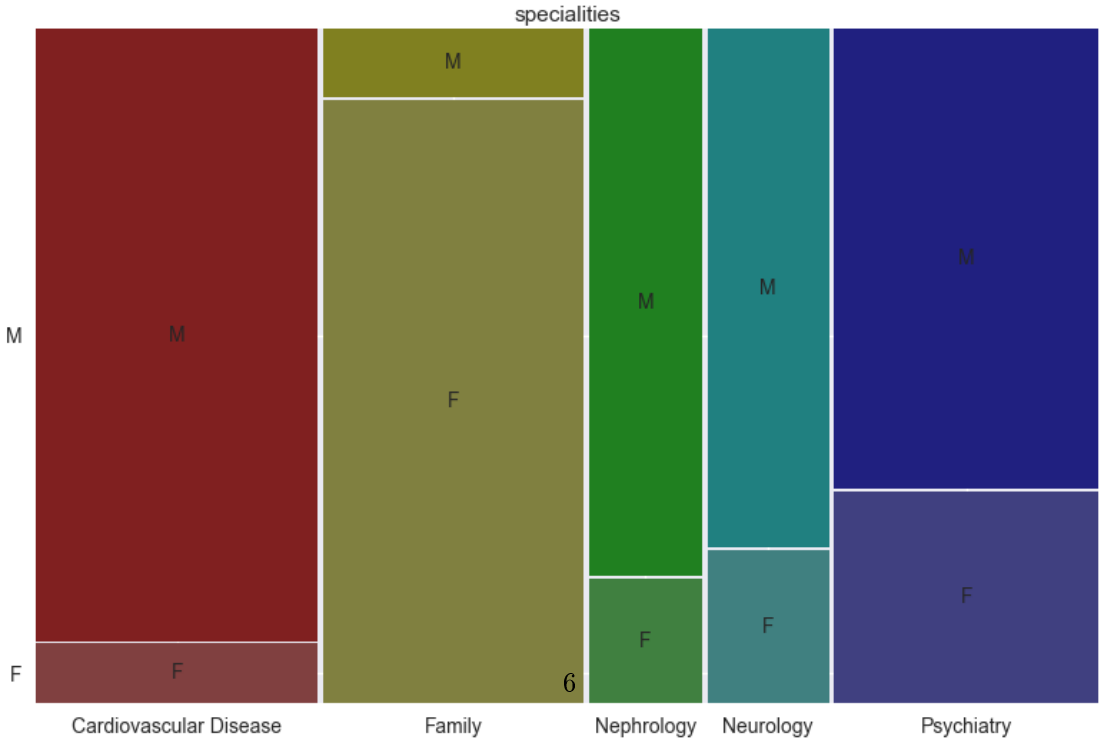


Figure 5: Region split, before (left) and after (right) removal



(a) Gender distribution according to years practicing



(b) Gender distribution according to specialties

Figure 6: Gender distribution according to various traits

Results

Note that the code used to obtain these results is well documented and contains further explanations.

Prediction of prescriber traits from prescriptions

Here are the prediction results, compared to a dummy model which randomly chooses a classification with uniform probability:

	Gender	Region	Specialty
RF	72%	50%	72%
BRF	75%	55%	74%
ARF	75%	57%	73%
LR	76%	66%	76%
Dummy	50%	25%	2.8%

Where LR is Logistic Regression, RF is Random Forest, BRF is Bagged Random Forest, and ARF is Adaboosted Random Forest. The corresponding confusion matrices can be found in Figures 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18.

ROC curves

Since the 'gender' field is a binary feature, we can extract ROC curves for it (which shows us the TP rate vs. FP rate of the model). In each graph the AUC (area under curve) is detailed on the bottom-right corner of the figure. The corresponding graphs can be found in Figures 19, 20, 21, 22. As we can see, in Logistic Regression the AUC is the highest (0.82).

Feature importance

The Random Forest classifier allows us to look at the importance it assigns to each feature. The higher the importance, the more the classifier uses this feature allows to separate the data into it's classes.

Let's look at the top 3 features that the Random Forest classifier assigned in each classification task:

Gender

1. OMEPRAZOLE - 0.018910
2. CITALOPRAM HBR - 0.010951
3. TRAZODONE HCL - 0.010797

Region

1. HYDROCODONE-ACETAMINOPHEN - 0.010411
2. GABAPENTIN - 0.009515
3. ATORVASTATIN CALCIUM - 0.009409

Specialty

1. VENLAFAXINE HCL ER - 0.015489
2. LISINOPRIL - 0.014903
3. LITHIUM CARBONATE - 0.013323

Modeling specialties according to prescriptions

The full results can be obtained by running our code.

Conclusions

In this project we explored prescription data in the United States, and showed that it is possible to predict certain traits of doctors using their prescription history. While predicting the doctor's specialty using his prescription might seem plausible (since it makes sense that different specialties have different medicines), predicting gender and region using these habits is more surprising. We saw that there are vast difference in gender splits between specialties, therefore it might be possible that our models use this fact in order to predict the gender (i.e, if the prescription is most likely related to a cardiovascular specialty, which is dominated by men, then the doctor is probably a man).

Furthermore, we saw that using LDA is also beneficial when trying to find out what medicines

'define' a specialty. This method can also be used as a starting point towards further work, as we will see in the next section.

Further work

Using the above information for real-world medicinal research, for example:

- Finding off-label medicines - are there certain medications officially developed for one purpose useful for certain specialties, but prescribed by doctors from other specialties? Using this information, unknown beneficial properties of medications could be discovered.
- Are there hidden predispositions among prescribes that led to the high accuracy of prediction of prescribe traits from prescriptions? The medical community can go over the results and research that more deeply.
- Using the 'intuition' about the feature importance from the Random Forest classifier to understand whether there is a prior disposition for different genders or regions. Maybe there are hidden factors that are not apparent in the data that influence prescriptions (for example, a medical firm in the south that incentives doctors to prescribe its medicines)?

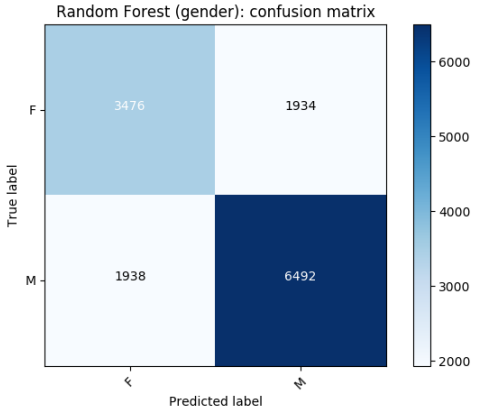


Figure 7: Gender Confusion Matrix (RF)

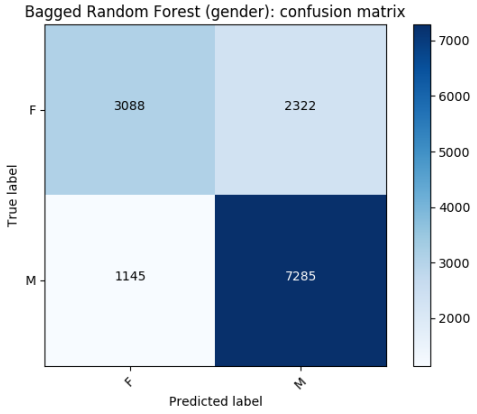


Figure 8: Gender Confusion Matrix (BRF)

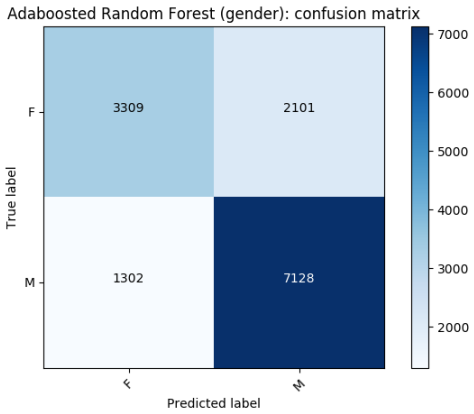


Figure 9: Gender Confusion Matrix (ARF)

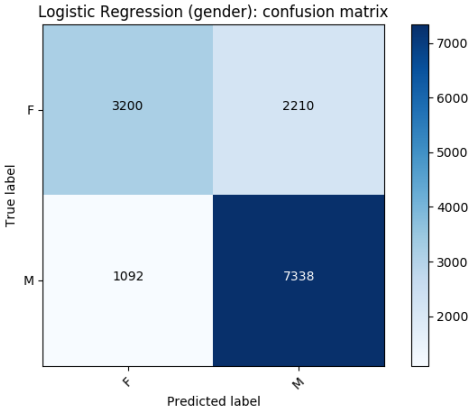


Figure 10: Gender Confusion Matrix (LR)

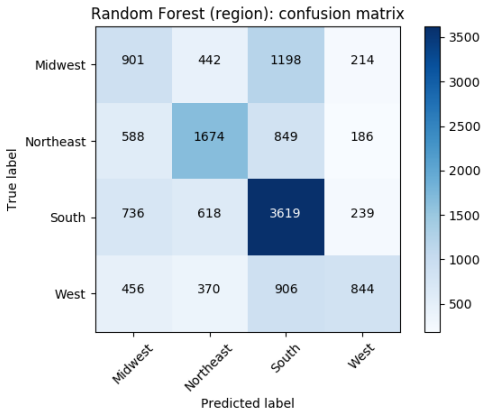


Figure 11: Region Confusion Matrix (RF)

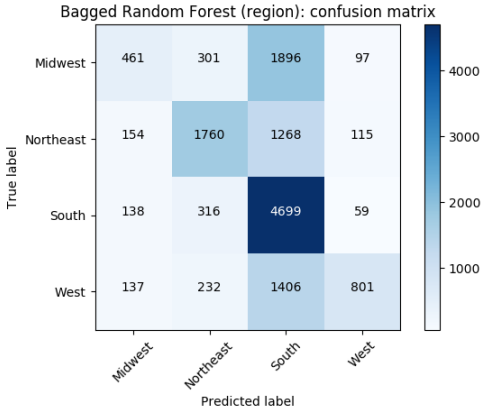


Figure 12: Region Confusion Matrix (BRF)

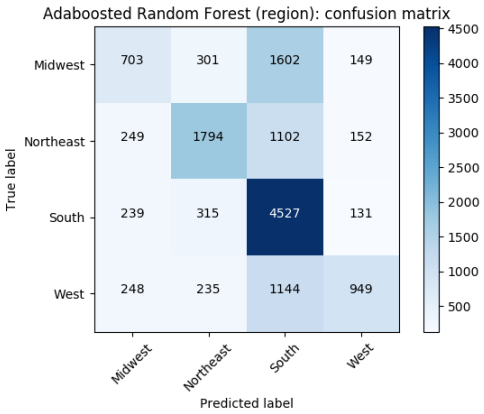


Figure 13: Region Confusion Matrix (ARF)

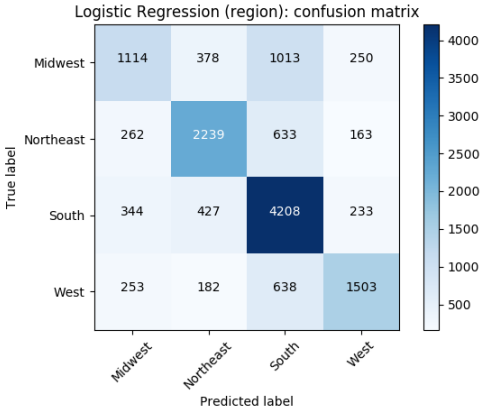


Figure 14: Region Confusion Matrix (LR)

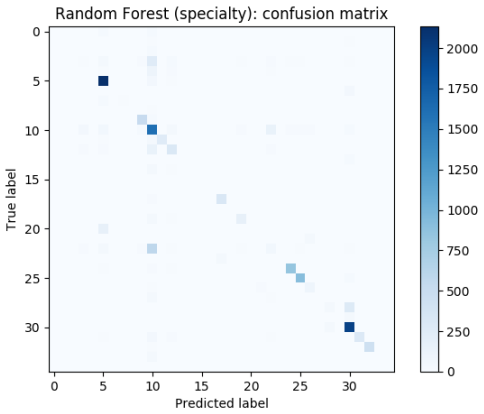


Figure 15: Specialty Confusion Matrix (RF)

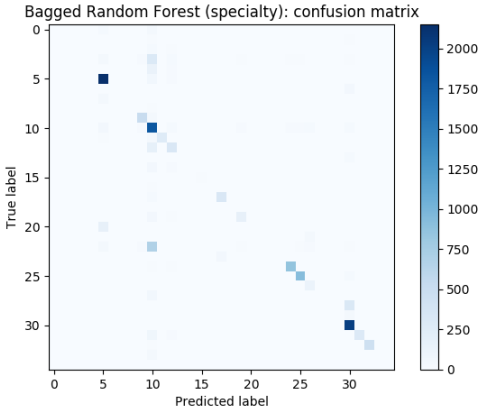


Figure 16: Specialty Confusion Matrix (BRF)

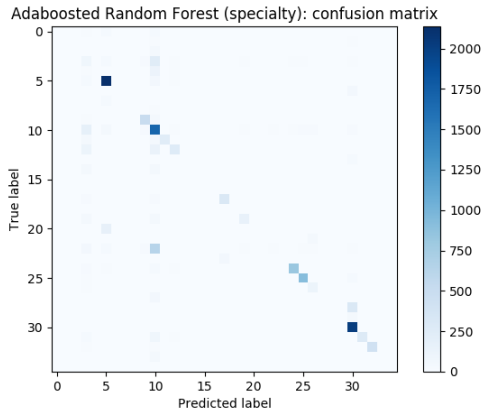


Figure 17: Region Confusion Matrix (BRF)

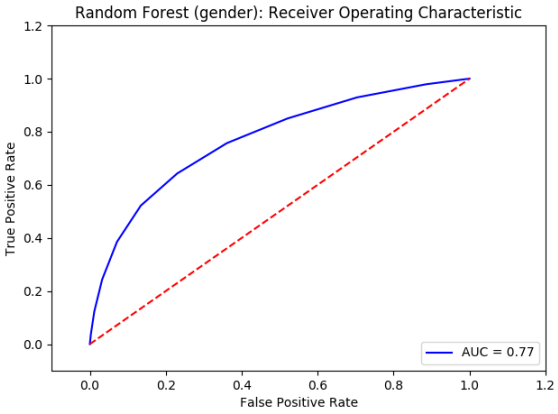


Figure 19: ROC Curve, Gender Classification (RF)

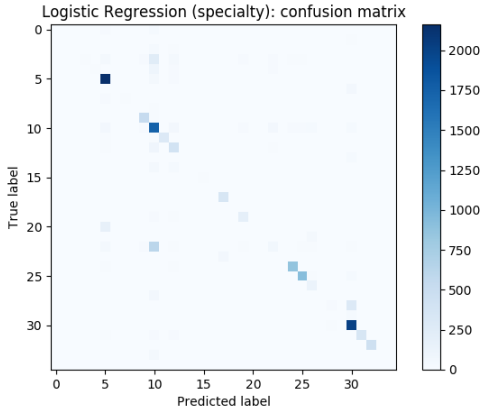


Figure 18: Specialty Confusion Matrix (LR)

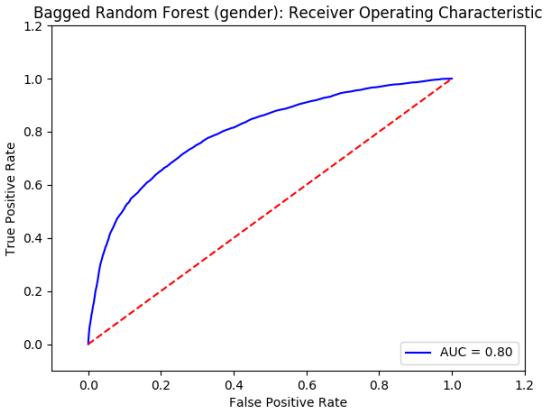


Figure 20: ROC Curve, Gender Classification (BRF)

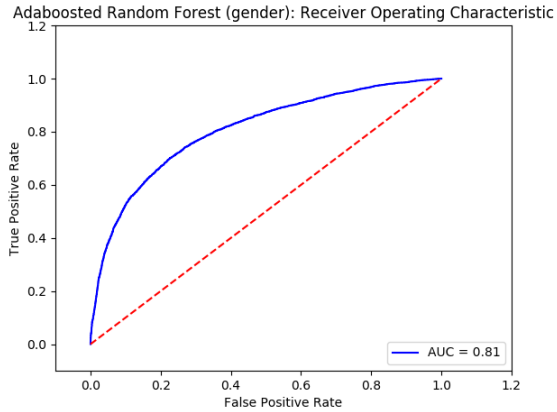


Figure 21: ROC Curve, Gender Classification (ARF)

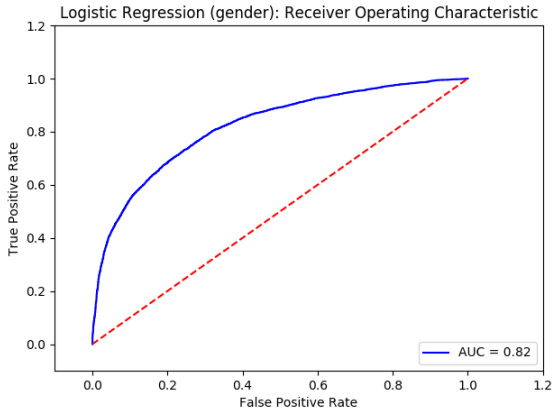


Figure 22: ROC Curve, Gender Classification (LR)