**Imputation of High-level Patient Racial Data using Patient Demographics, Diagnostics, and Clinical Discharge Note Embeddings**

**Background**

Missing demographic data, especially missing racial data within electronic health records, can amplify biases and reduce transparency on the disparities within the data. Furthermore, models or applications that leverage underlying healthcare data with missing patient demographics can carry the same biases. Racial and demographic data are commonly missing in EHR records, due to their self-voluntary nature, and are also often missing of datasets used for biomedical informatics.

The primary method of indirectly estimating race/ethnicity relies on residential address and surname

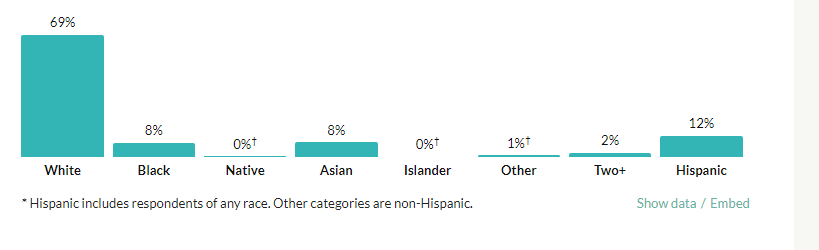
Information and is known as the Bayesian Improved Surname and Geocoding (BISG). Numerous studies augment BSIG with other patient data around insurance status, income, and embed more sophisticated location analysis. However, in numerous contexts, only anonymized data or data where such information is redacted is available.

The primary aim of this study is to explore an alternative approach towards indirectly estimating race that takes the health records in the form unstructured clinical notes, patient diagnostics, and what limited other demographics are available in order. While these models may not offer a substitute to BSIG, they may be augmented with BSIG-style techniques to enhance the quality of racial prediction or be the foundation for imputation in resource-constrained environments.

**Dataset and Task**

Specifically, we look at the MIMIC-III Dataset, analyzing the distribution of race. Within the MIMIC-III dataset, we observe a systemic undercounting of Black, Non-white Hispanic, Asian, and Other (Multiracial, American Indian) communities, and higher representation of White patients.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RACE** | **MIMIC Patient Counts** | **Proportion** | **Real**  **(Census)** | **Differential** |
| WHITE | 41368 | 77.9% | 69.9% | +8.0% |
| BLACK | 5785 | 10.9% | 13.4% | -2.5% |
| HISPANIC | 2136 | 5.2% | 8.7% | -3.5% |
| ASIAN / PACIFIC ISLANDER | 2025 | 3.8% | 6.1% | -2.3% |
| OTHER | 1766 | 3.3% | 3.9% | -0.6% |
| UNKNOWN | 5896 | N/A | N/A | N/A |



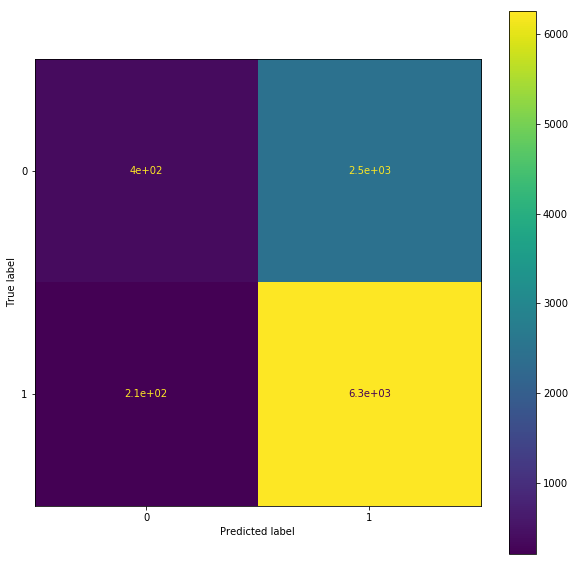
Additionally, we observe significant differences in the number of notes dependent on the race of the patient, with just the median and modes across a large number >1.5K patients in each case.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **RACE** | **MIMIC Notes**  **(Median)** | **MIMIC Notes**  **(Mean)** | **MIMIC Notes**  **(Total)** | **Proportion** | **Real** | **Differential** |
| WHITE | 29.0 | 66.1 | 2718852 | 73.2% | 69.9% | +3.3% |
| BLACK | 39.0 | 110.3 | 635688 | 17.1% | 13.4% | +3.7% |
| HISPANIC | 31.0 | 70.1 | 148727 | 4.0% | 8.7% | -4.7% |
| ASIAN / PACIFIC ISLANDER | 17.0 | 49.2 | 98250 | 2.6% | 6.1% | -3.5% |
| OTHER | 24.0 | 63.2 | 110696 | 3.0% | 3.9% | -0.9% |
| UNKNOWN | 19.0 | 41.2 | 239917 | N/A | N/A | N/A |

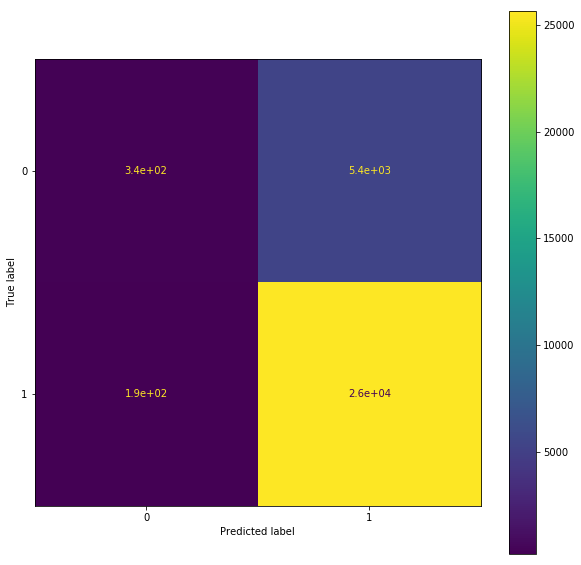
For simplicity, the task of multi-class racial classification is reduced to a binary classification of white versus non-white. Note, that the dataset is significantly unbalanced towards white patients – a bias that propagates through the models trained, and the smaller proportion of non-white patients inhibits potential multi-class training.

**Feature Selection and Classification Models**

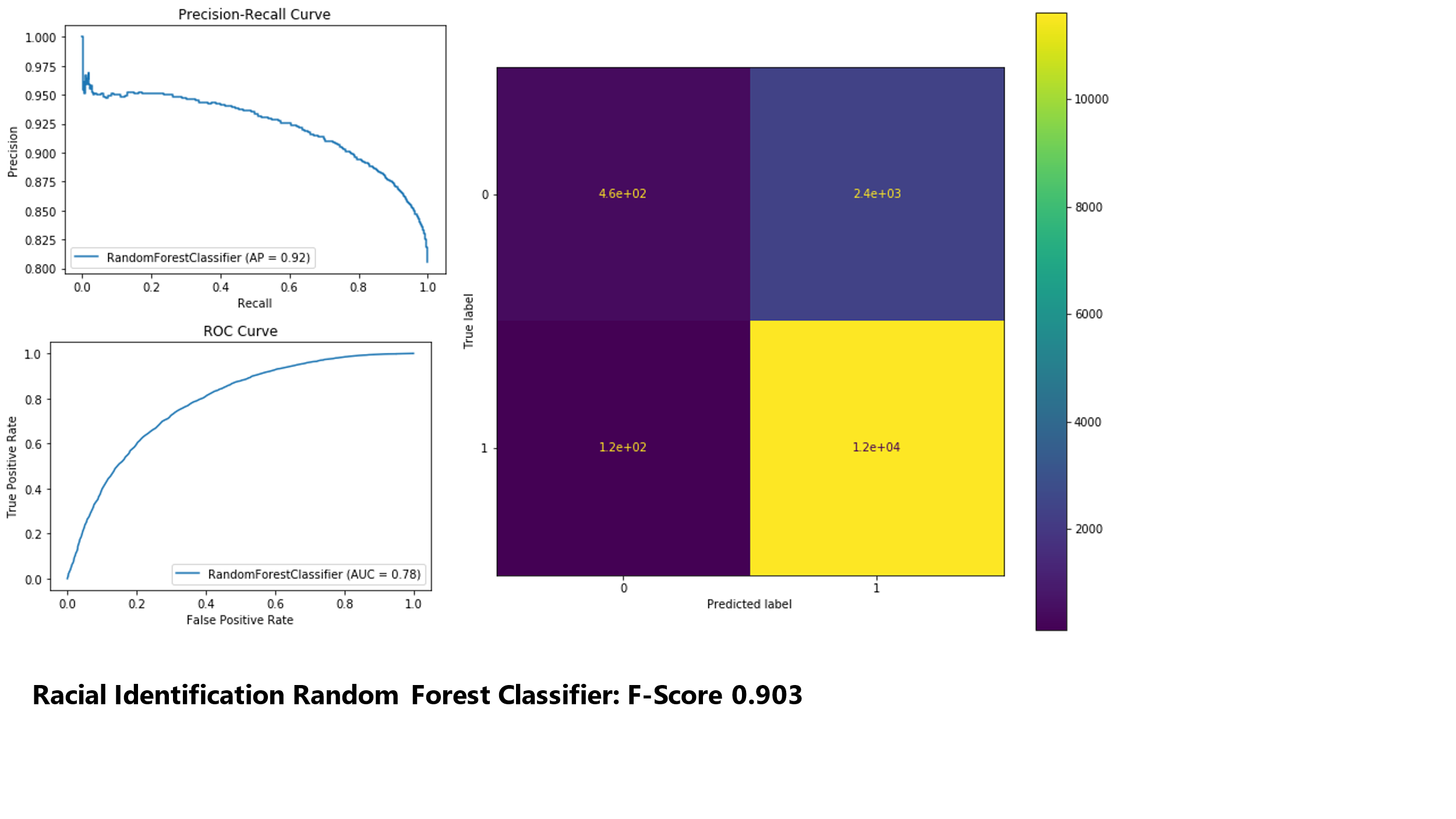
First, a set of the patients’ non-race demographics are selected as features. These other demographics include their insurance, marital, language, and religion status. Note, that a number of these demographics are reported as unknown and are collected voluntarily from the patient. A Random Forest Classifier trained with these core demographics alone achieves a F-score of ~0.78-0.82, and is the basis of the classification model.

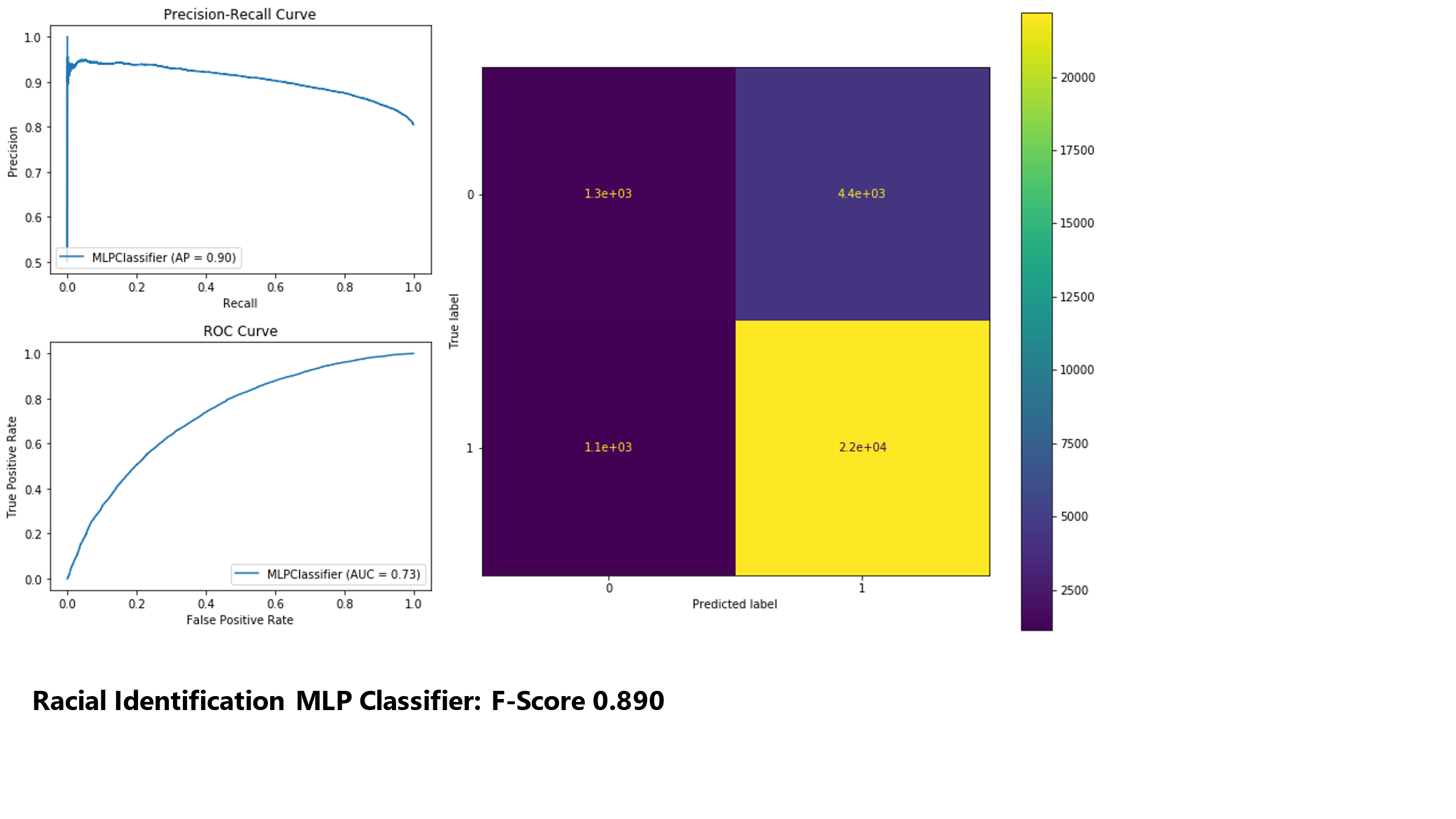


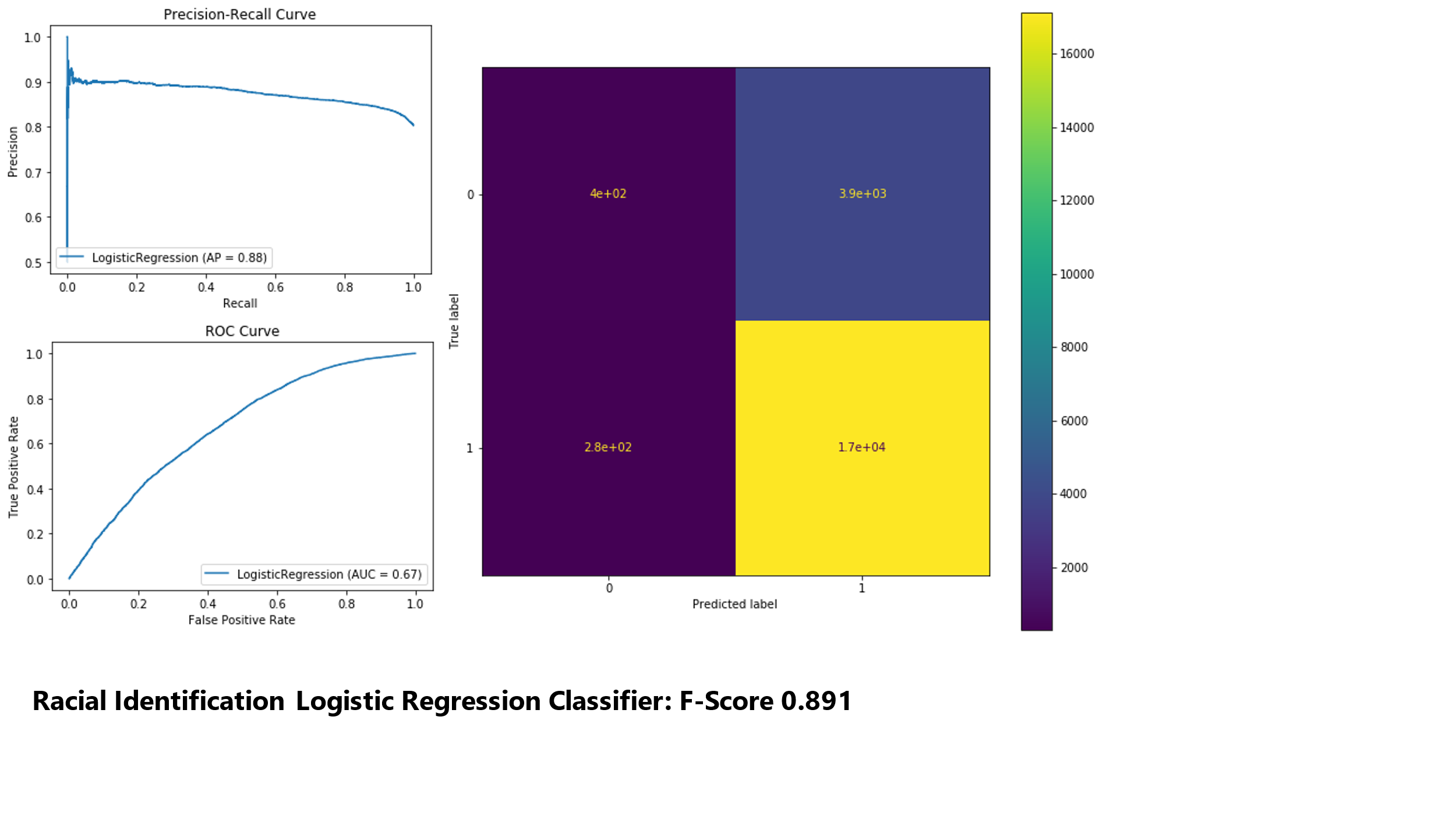
Next, the ICD code statuses of the patients indicating their medical history are concatenated to the other demographics. This is associated with a sizable improvement in F-score to nearly 0.90.



Finally, TF-IDF Vectorization of the most recent clinical discharge note of each patient was produced with the top 5,000 tokens by frequency in text. Only tokens correlated with text above 0.04 were retained as features. This is concatenated to existing features along with the counts of the different note types associated with each patient. This is associated with modest gains in the model’s performance, but improved performance in identifying non-white patients. Note, that the choice of TF-IDF is deliberate, with its performance better than Doc2Vec embeddings of the entire discharge note. Additionally, clinical discharge notes are selected as the note type of focus because they have the second highest prevalence of racial key words (to social history), and are widely available (less than 5,000 patients have social history notes, making the limited size of the training data damaging to model performance).







Note that these results, and subsequent results are selected from one of five-fold cross validation, with similar results in other validation splits with the same data.

The performance of the Random Forest Classifier, achieving an AUC of 0.78, is significantly higher than the 0.5-0.6 AUC achieved in a similar study of imputing basic demographics (gender, age) with embeddings of notes, specifically the study, ‘What’s in a Note’, where BOW’s, LSTM, and Embeddings are used to capture relevant aspects of discharge notes for the same task of white versus non-white ethnicity classification with the MIMIC-III dataset. This can be attributed to the different set of demographics used, the more sophisticated feature selection from notes, ad usage of diagnostic codes.

**Imputation of Marital Status**

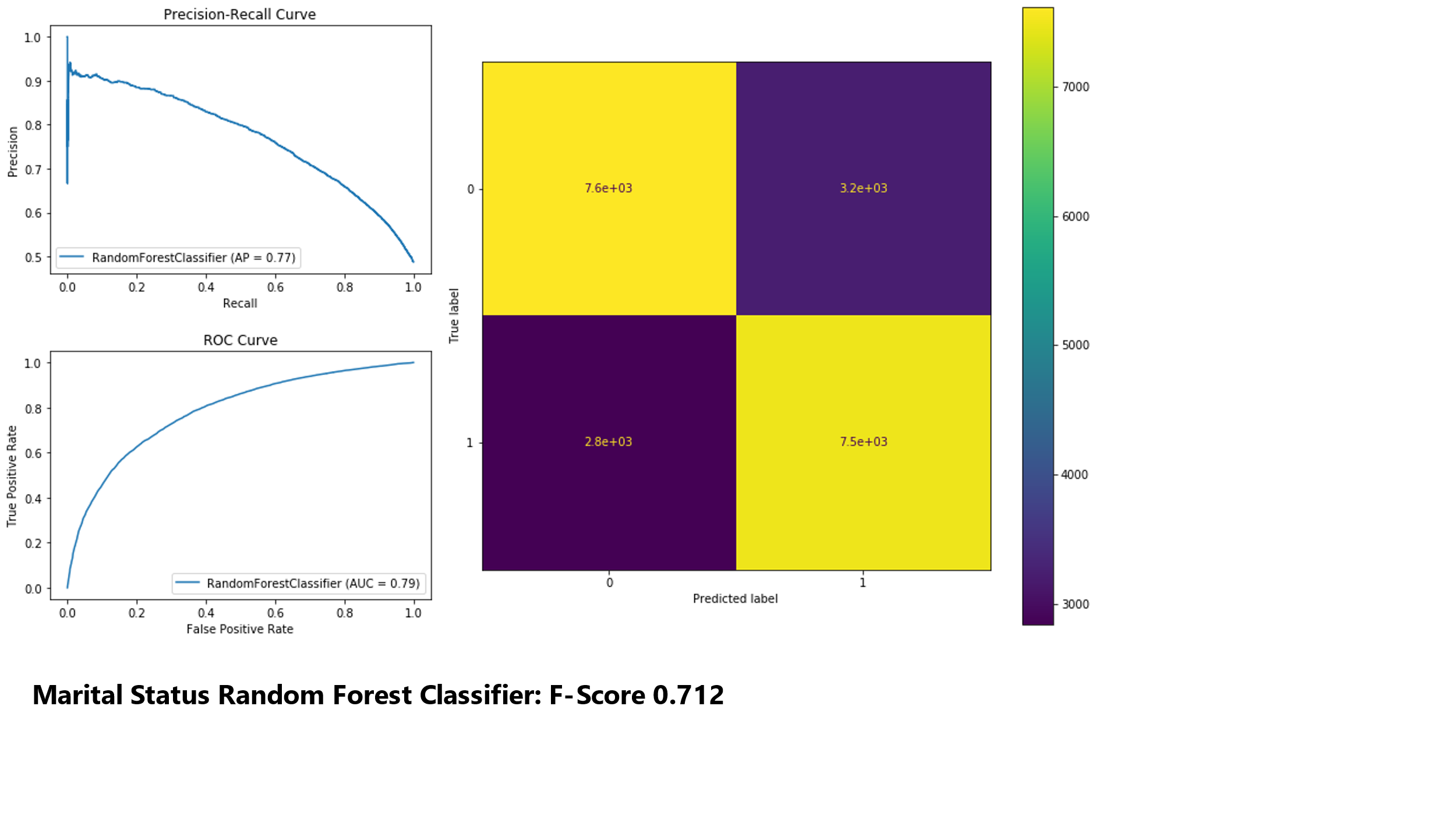
Another patient demographic where a large proportion is unknown is the marital status of patients. Because of inconsistencies associated with changes in marital status and the highly voluntary nature of reporting marital status (compared to medical history or insurance status), it is often missing or out of date in patient records.

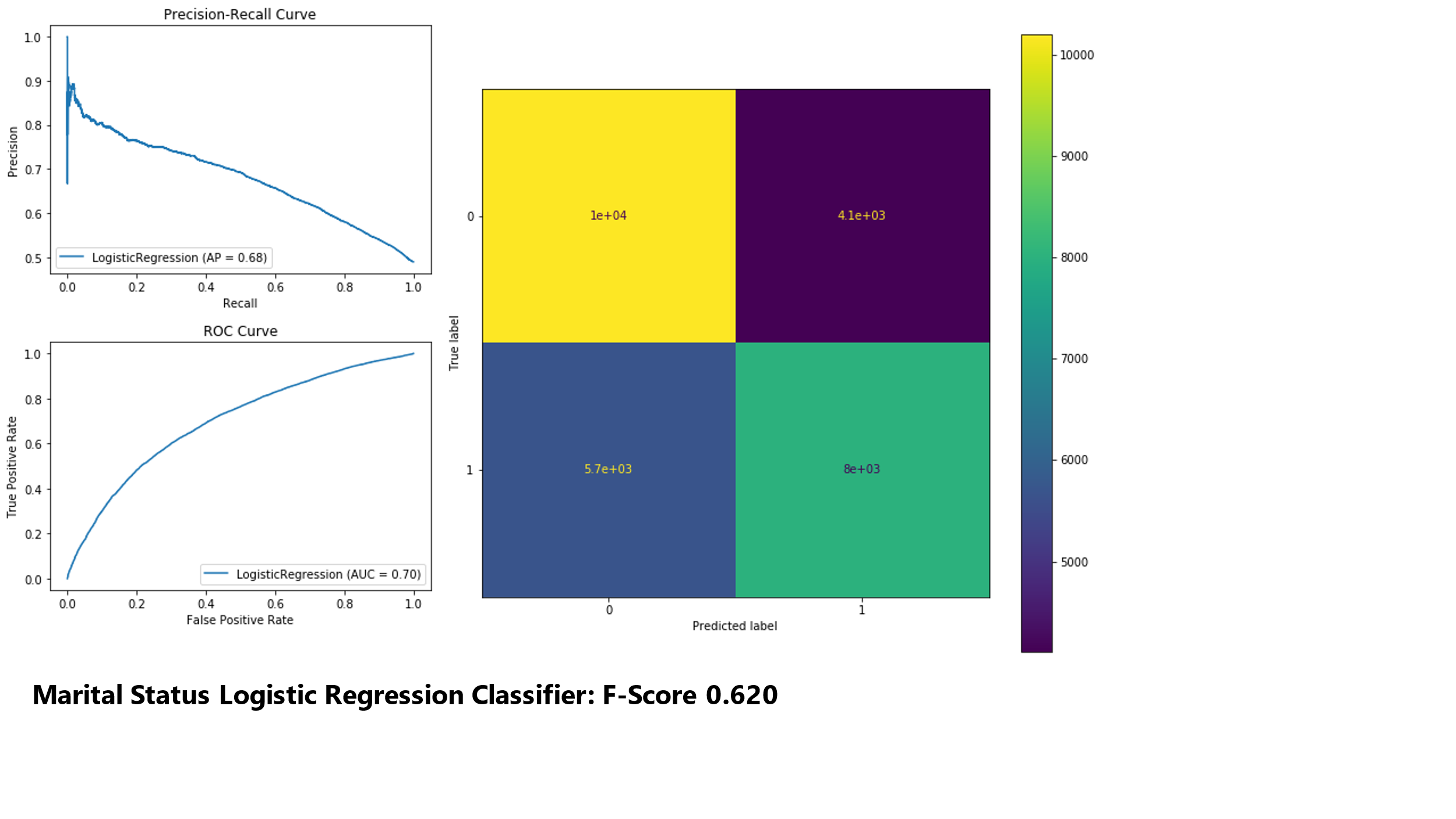
|  |  |
| --- | --- |
| **Marital Status** | **Raw Count** |
| Married | 24239 |
| Single | 13254 |
| Widowed | 7211 |
| Divorced | 3213 |
| Separated | 571 |
| Unknown | 345 |
| Life Partner | 15 |

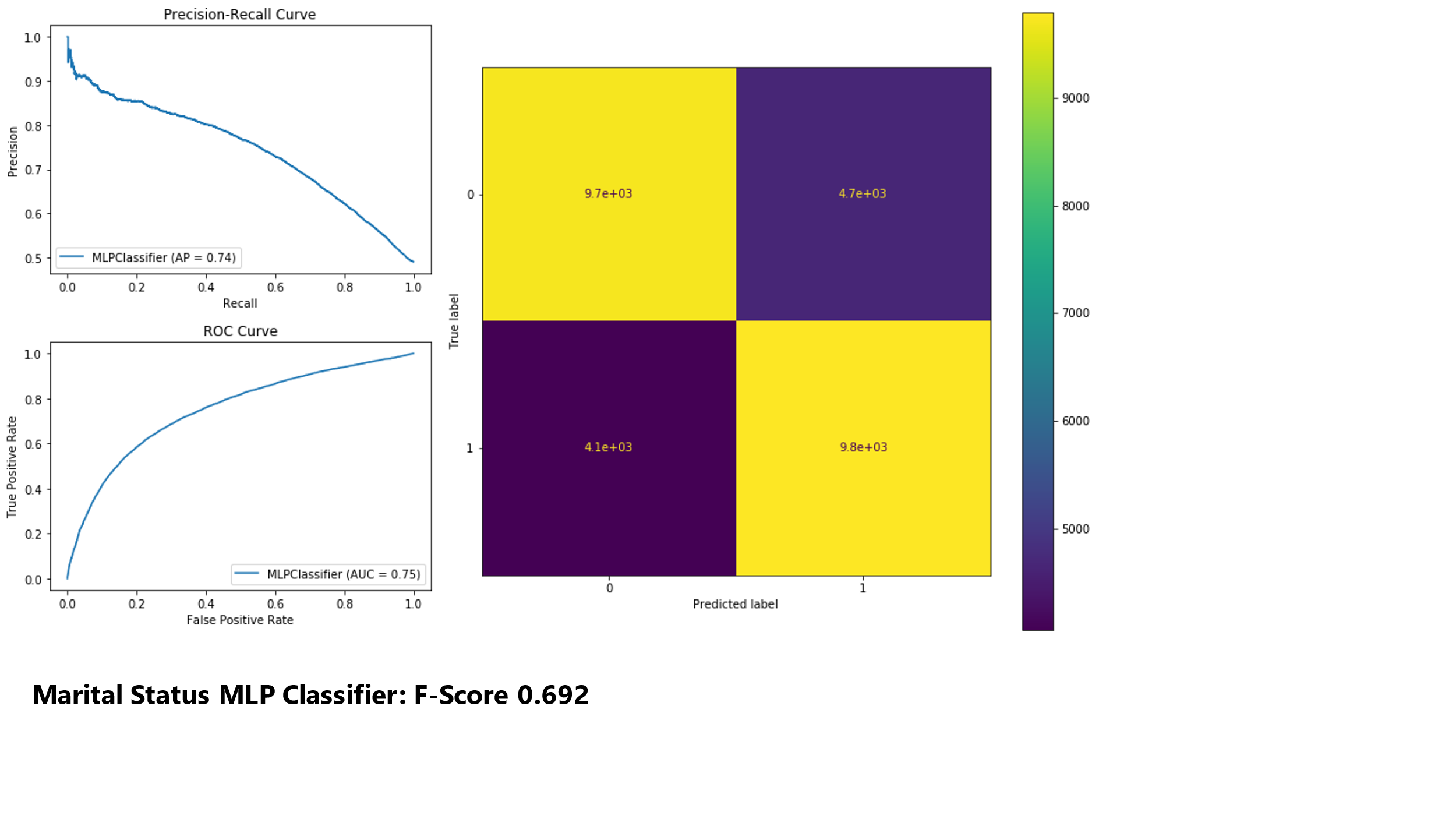
After removing duplicates, and updating to the last marital status record per subject, we simplify to a binary task of separating individuals identified as married from not-married (i.e., single, windowed, divorced, separated). Life partner is included with married, and unknown is excluded.

|  |  |
| --- | --- |
| **Marital Status (Binary)** | **Count** |
| Married (Life Partner included) | 17201 |
| Not Married | 17991 |

The Random Forest classifier achieves an AUC of 0.79, indicating moderate success at the classification task. While performance by the MLP classifier is similar, performance by the Logistic Regression classifier is considerably worse.





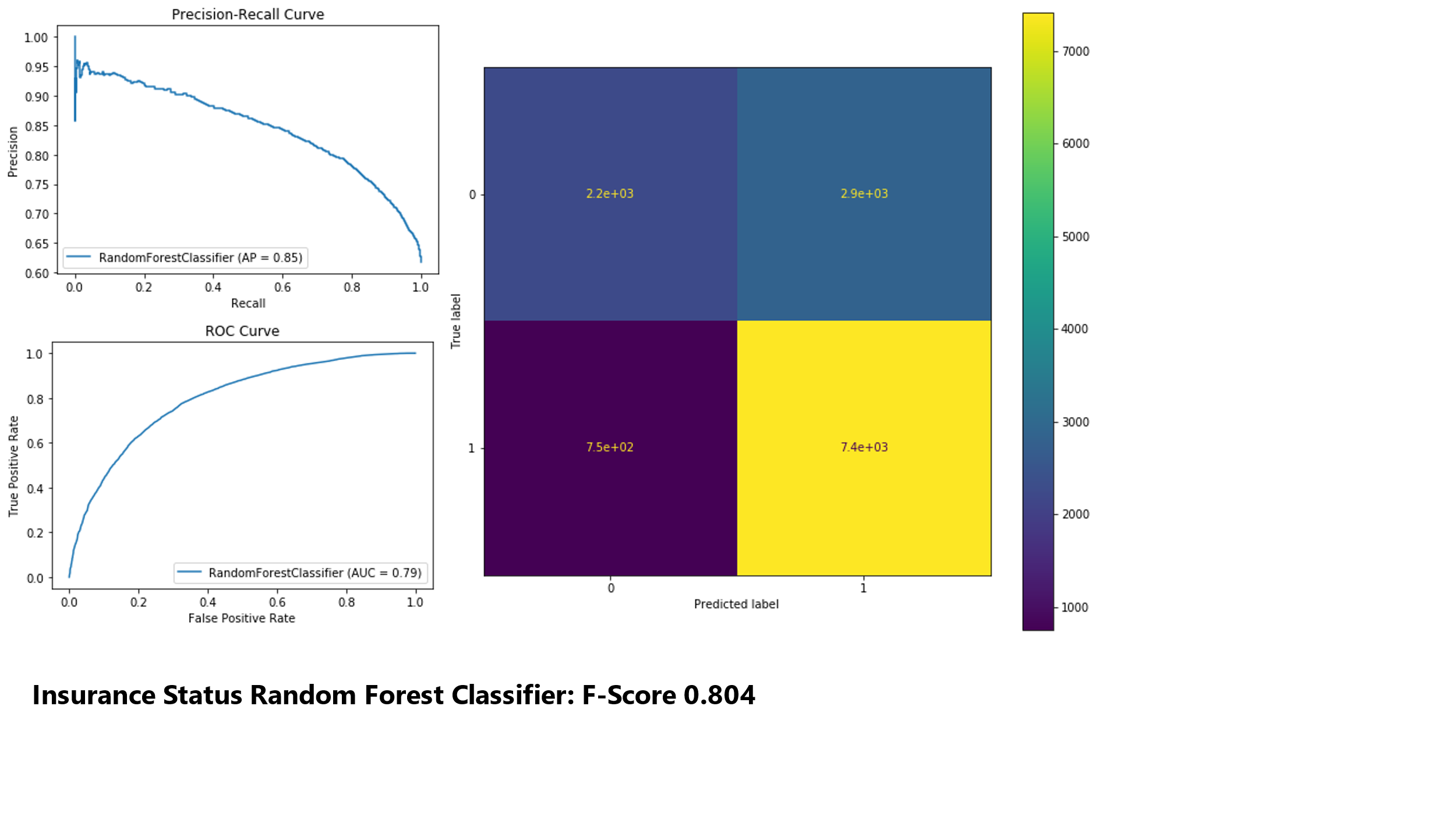
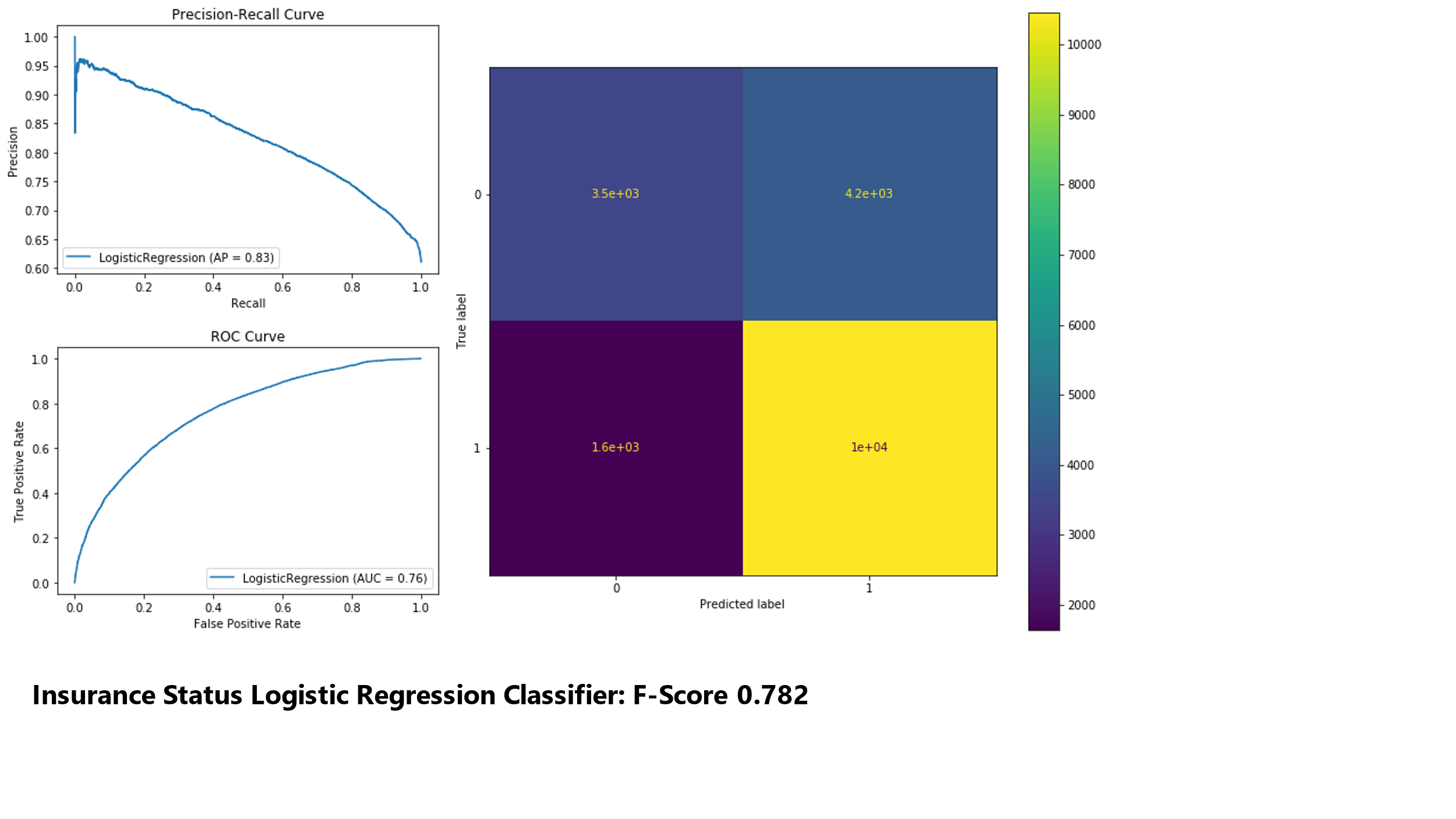
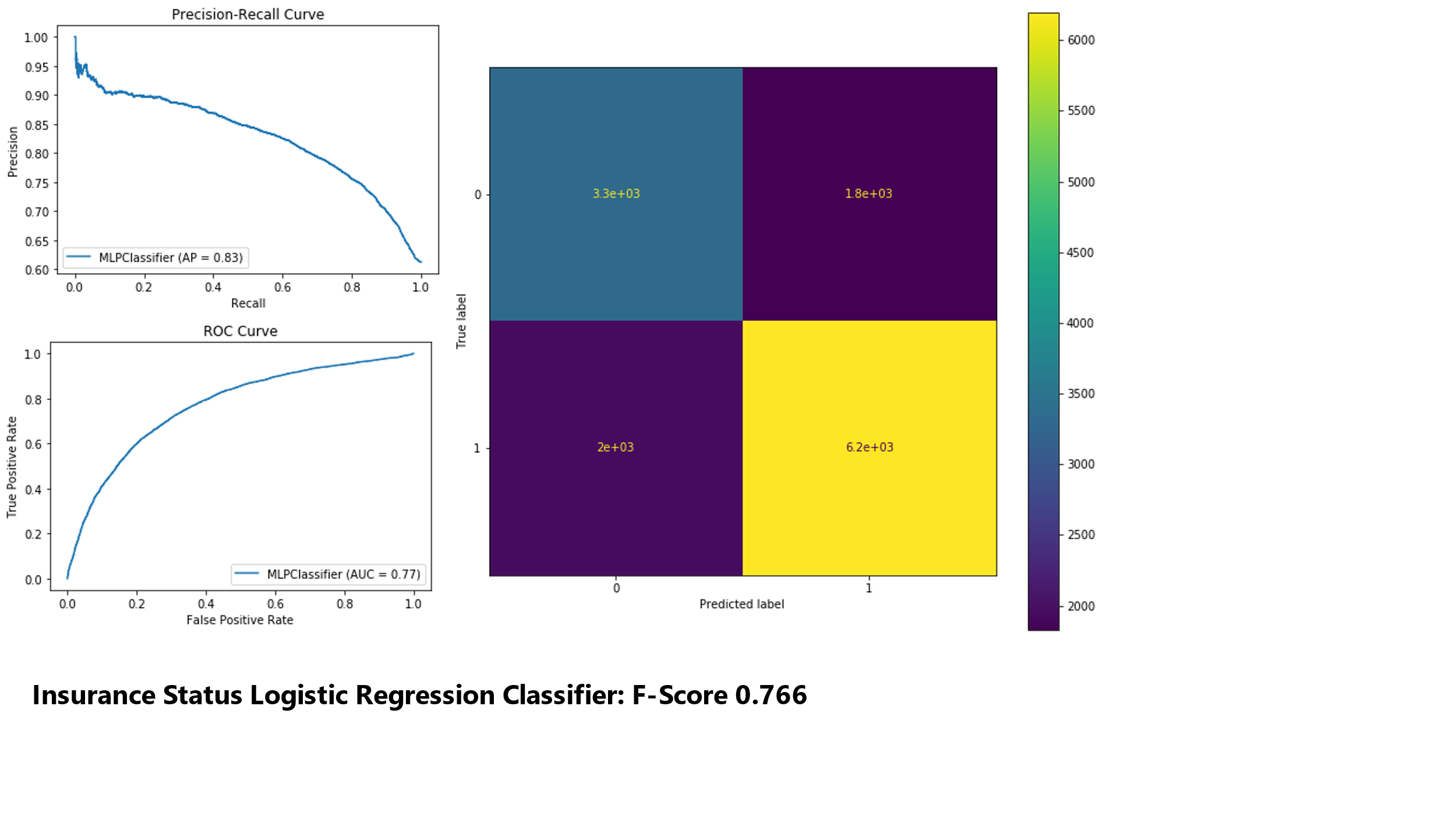


**Imputation of Insurance Status**

In a similar style to marital status classification, insurance status prediction is converted to binary classification of public versus private.

|  |  |  |
| --- | --- | --- |
| **Insurance Status** | **Raw Count** | **Binary Status** |
| Medicare | 28215 | Public |
| Private | 22582 | Private |
| Medicaid | 5785 | Public |
| Government | 1783 | Public |
| Self-Pay | 611 | Private |

The top performing model is again the Random Forest Classifier, achieving a F-score of above 0.8 and AP of 0.85. The moderate success indicates the model can be extended beyond application of racial demographics to tasks such as marital status and insurance status prediction.

**Future Work**

The actual embedding of clinical notes remains an area needing of improvement. Using alternative document embeddings such as BERT or GloVe or a predefined set of racial key words may be viable options. Second, the task can be extended from binary classification to multi-class classification, with analysis of performance at a more granular level (ex. Black versus Non-Black, Hispanic versus Non-Hispanic in addition to White versus Non-White). Finally, multiple formats of notes may be concatenated at a patient level (ex. social history + discharge) in order to maximize the amount of retained records per patient instead of arbitrary exclusion.

**References**

**MIMIC-III, a freely accessible critical care database.** Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016). DOI: 10.1038/sdata.2016.35.

Haas A, Elliott MN, Dembosky JW, et al. **Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity.** Health Serv Res. 2019;54:13–23. <https://doi.org/10.1111/1475-6773.13099/>

Boag, Willie et al. **“What's in a Note? Unpacking Predictive Value in Clinical Note Representations.”** AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science vol. 2017 26-34. 18 May. 2018

Grundmeier RW, Song L, Ramos MJ, Fiks AG, Elliott MN, Fremont A, Pace W, Wasserman RC, Localio R. **Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data.** Health Serv Res. 2015 Aug; 50(4): 946-60. https://doi.org/10.1111/1475-6773.12295/

Le, Quoc; Mikolov, Tomas;, **Distributed Representations of Sentences and Documents;** Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2):1188-1196, 2014.

Kim JS, Gao X, Rzhetsky A. **RIDDLE: Race and ethnicity Imputation from Disease history with Deep Learning.** PLoS Comput Biol. 2018 Apr 26; 14(4) <https://doi.org/10.1371/journal.pcbi.1006106/>

Xue Y, Harel O, Aseltine RH Jr. **Imputing race and ethnic information in administrative health data.** Health Serv Res. 2019 Aug;54(4):957-963. doi: 10.1111/1475-6773.13171. Epub 2019 May 17. PMID: 31099021; PMCID: PMC6606547.

Long JA, Bamba MI, Ling B, Shea JA. **Missing race/ethnicity data in Veterans Health Administration based disparities research: a systematic review.** J Health Care Poor Underserved. 2006 Feb;17(1):128-40. doi: 10.1353/hpu.2006.0029. PMID: 16520522.

Fremont, Allen et al. **“When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities.”** Rand health quarterly vol. 6,1 16. 20 Jun. 2016

Ryan R, Vernon S, Lawrence G, Wilson S. **Use of name recognition software, census data and multiple imputation to predict missing data on ethnicity: application to cancer registry records.** BMC Med Inform Decis Mak. 2012 Jan 23;12:3. doi: 10.1186/1472-6947-12-3. PMID: 22269985; PMCID: PMC3353229.

Adjaye‐Gbewonyo D, Bednarczyk RA, Davis RL, Omer SB. **Using the Bayesian improved surname geocoding method (bisg) to create a working classification of race and ethnicity in a diverse managed care population: a validation study.** Health Serv Res. 2014;49(1):268‐283.

Storey P, Murchison AP, Dai Y, et al. **Comparing methodologies for imputing ethnicity in an urban ophthalmology clinic.** Ophthalmic Epidemiol. 2014;21(2):106‐110.