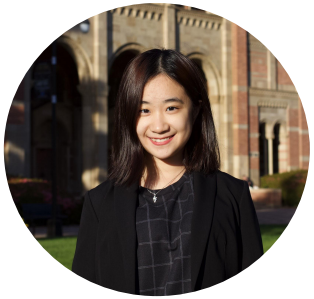
Persian Versus Caucasian Eyes

Team Members



Jinyi Li

Statistics

Jiancong Qi

Statistics

Joy Wang

Statistics,
Economics

Zheng Wang

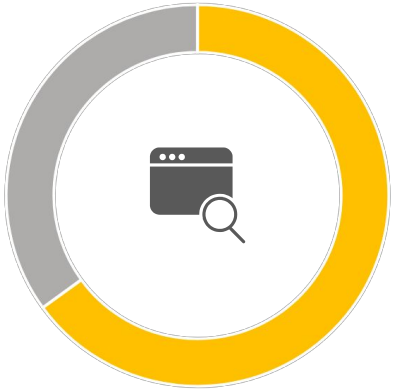
Statistics,
Computational Math

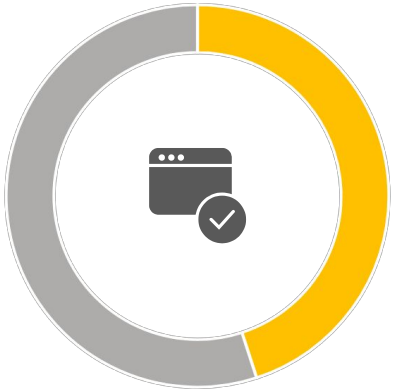
Wenxin Zhou

Statistics,
Computational Math

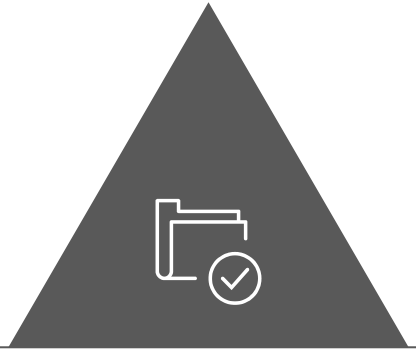
Problem Statement

- Can we use features to predict whether an eye is Persian?
- Can being Persian predict certain measurements about eye features?





Data Cleaning



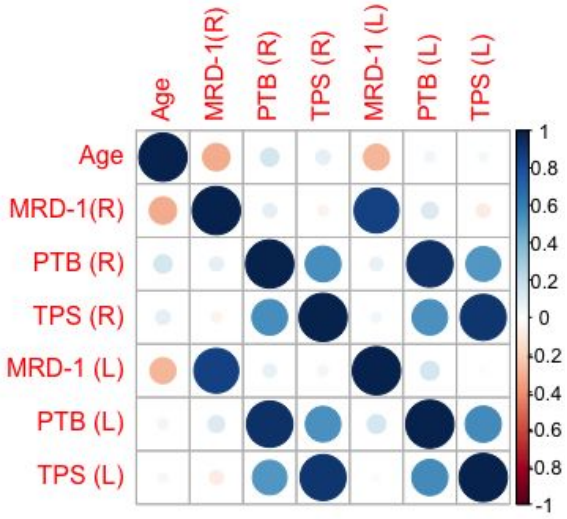
Build model without the online datasheet

Built model separately for male and female

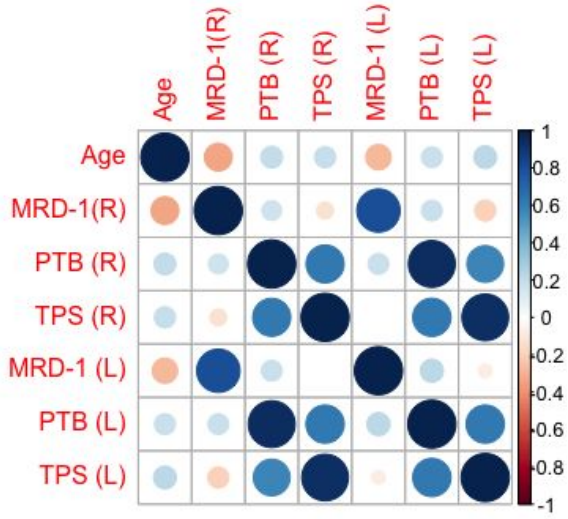
Exploratory Data Analysis

- The predictors we used are MRD-1 (L),MRD-1(R),PTB (R), TPS (R), TPS (L), PTB (L) and we want to find the correlation between these features.

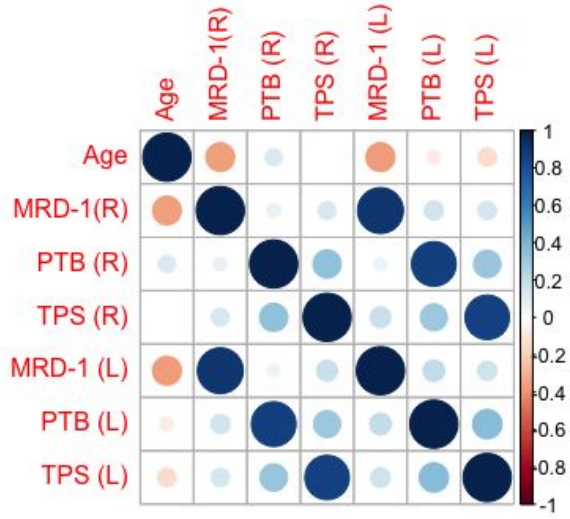
- Correlation plots for Caucasian, Persian and All people:



All



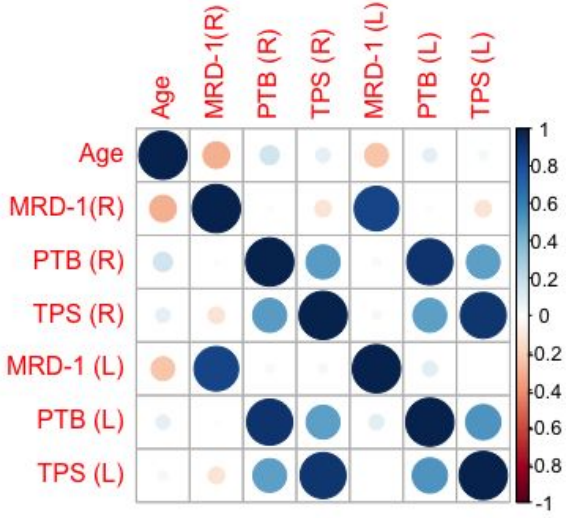
Caucasian



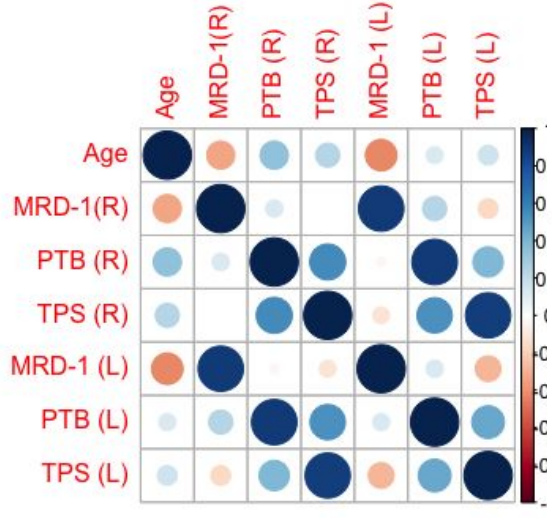
Persian

Exploratory Data Analysis

- Correlation plots for Female and Male
- Investigate linear relation between age and MRD

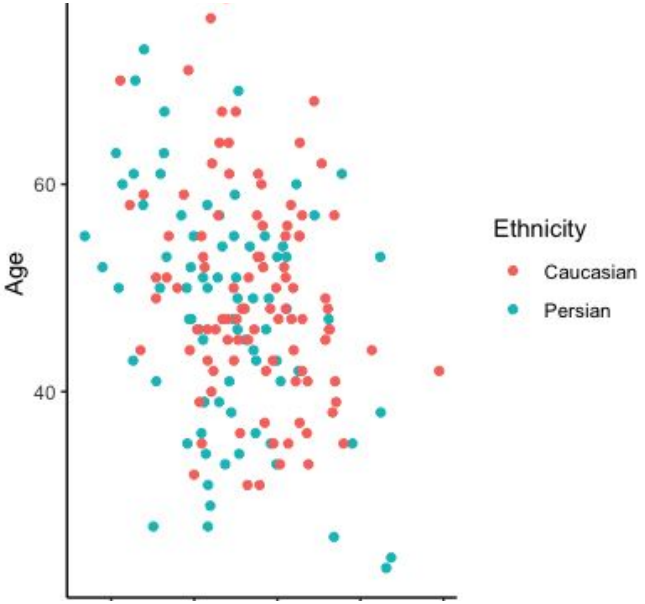


Female



Male

MRD



Exploratory Data Analysis

Correlation Analysis:

Observation:

1. MRD, PTB and TPS all show fairly high correlation between left and right eyes (~0.8).
2. TPS and PTB correlate with each other (~0.6). Such correlation is slightly higher among Caucasians and females.
3. Although not very strong, age negatively correlates with MRD, particularly among males (~0.4).

Insight:

1. We will only use features of one side of eyes in models to avoid multicollinearity.
2. We should consider adding age to the model as a covariate.

Note: The gender distribution of the dataset is very skewed, with 145 observations of females and only 34 observations of males. Therefore, the conclusion we got from the male samples could be biased.

Exploratory Data Analysis

Distribution Analysis

Observation

1. MRD shows Gaussian distribution, while TPS seems lognormal. PTB is ambiguous and may be either.
2. MRD is apparently differentiated by ethnicity.
3. After splitting the data by gender, both PTB and TPS show obvious differences between the ethnic groups.
4. 2. The distribution of PTB(L) for Male looks abnormal. Since PTB(R), which is highly correlated to PTB(L), shows a normal Gaussian distribution, we assume that the exotic shape arises due to the small sample size (only 21 Caucasians and 13 Persians in this sample).

Insight

1. All of the three eye features, MRD, TPS and PTB, could be significant to predict ethnicity.
2. The models might perform better if we log-transform TPS.
3. We should create separate models for both genders instead of neglecting the covariate

Exploratory Data Analysis

Investigation of Derived Variables

Observation

1. According to the distribution plots, the derived variables, MRD/TPS and MRD/PTB are significantly differentiated by gender and negatively correlate with age.

Insight

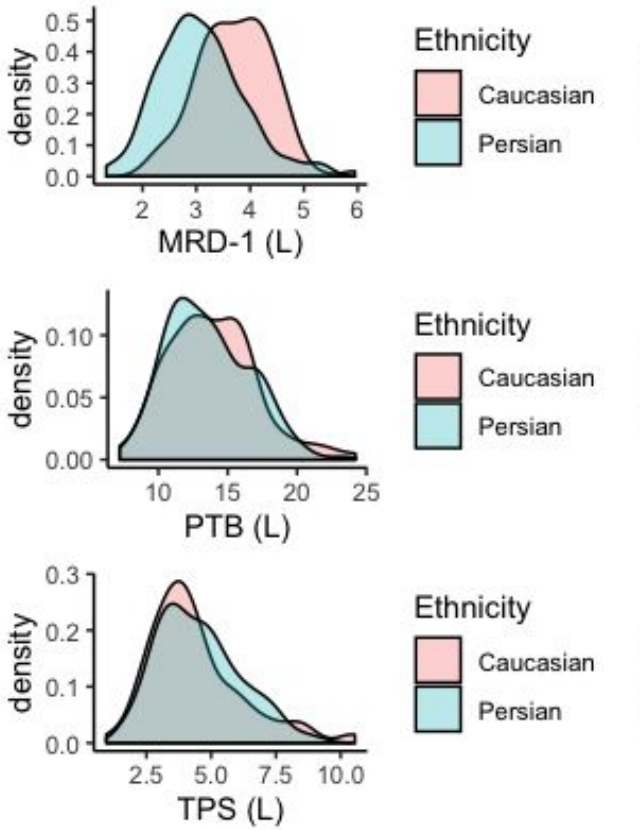
1. The above two observations suggest that we can potentially predict the two derived variables, MRD/TPS and MRD/PTB, with gender and age.

Note: Since MRD is a significant feature itself, it is probable that the significance of MRD/TPS and MRD/PTB comes from it, which undermines the value of this observation.

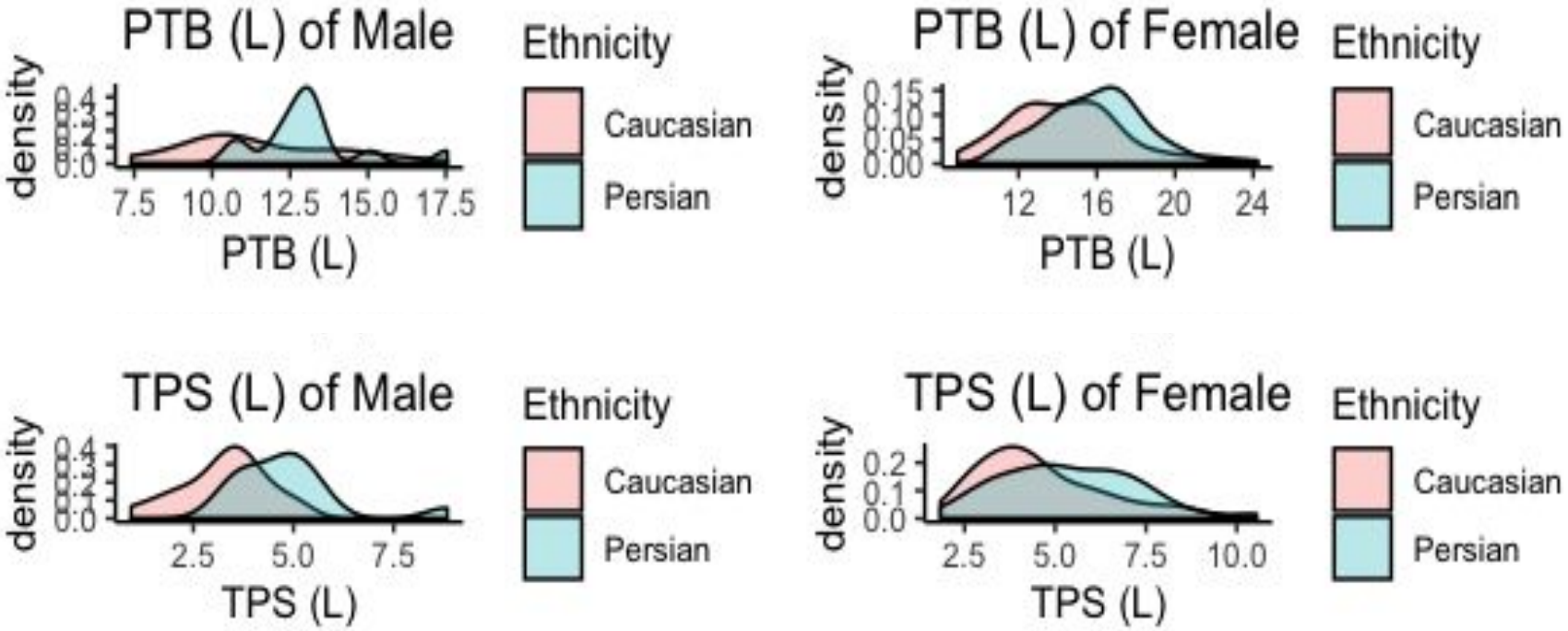
Distribution

Based on the density distribution of each eye features we can find:

1. MRD shows Gaussian distribution, while TPS seems lognormal. PTB is ambiguous and may be either.
2. Caucasian and Persian have obviously different MRD on both left and right eyes.



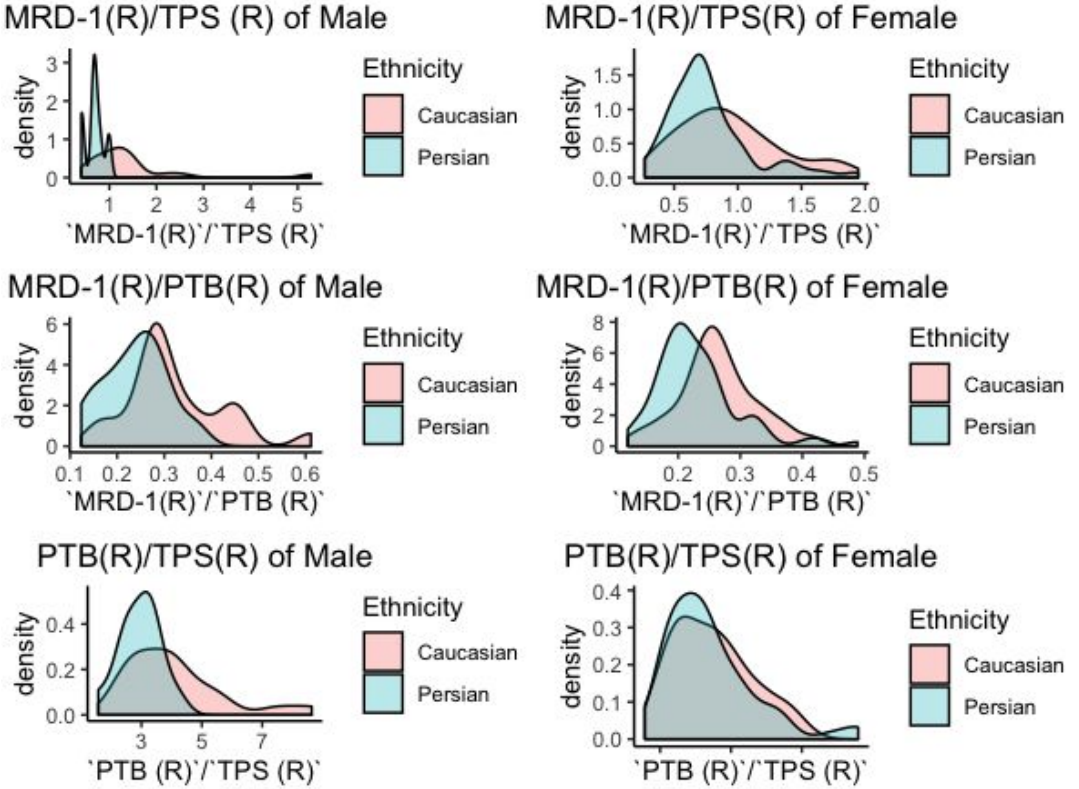
Distribution, partitioned by gender



Distribution of Variables Derived from Eye Features, partitioned by gender

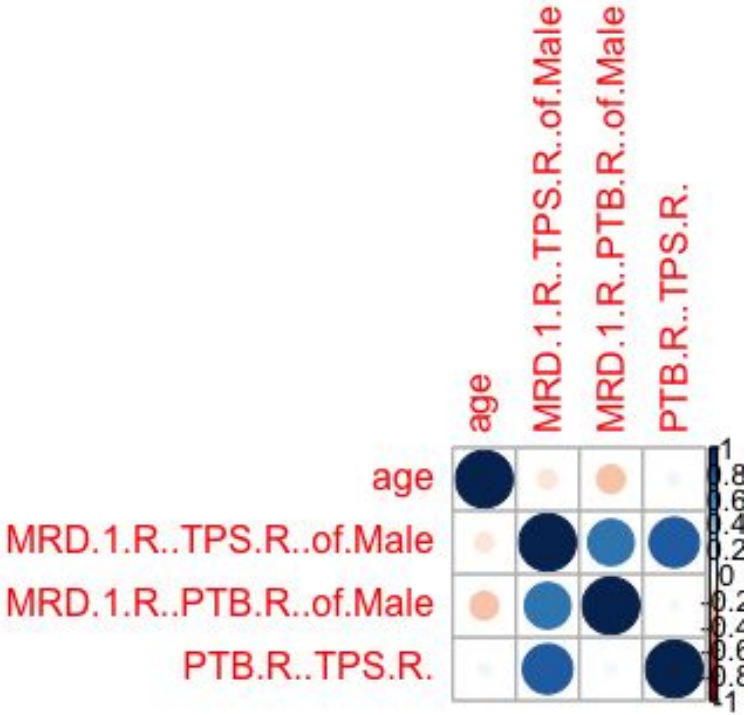
Based on the density distribution of variables derived from eye features we can find:

The derived variables, MRD/TPS and MRD/PTB are significantly differentiated by gender.

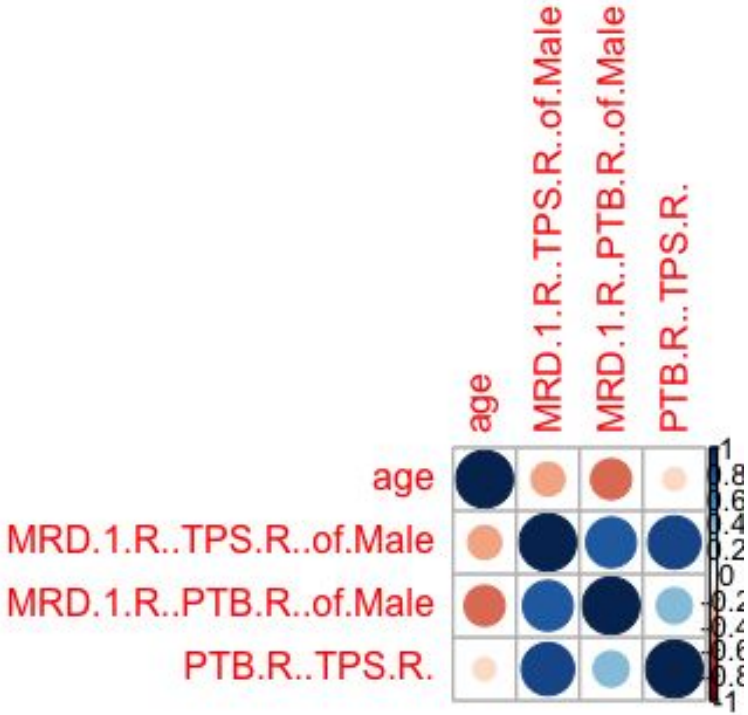


Correlation plots

- Males' ages correlate with their MRD/TPS and MRD/PTB, with coefficients of 0.34 and 0.5 respectively. Such correlations are weaker on females, but they still exist.

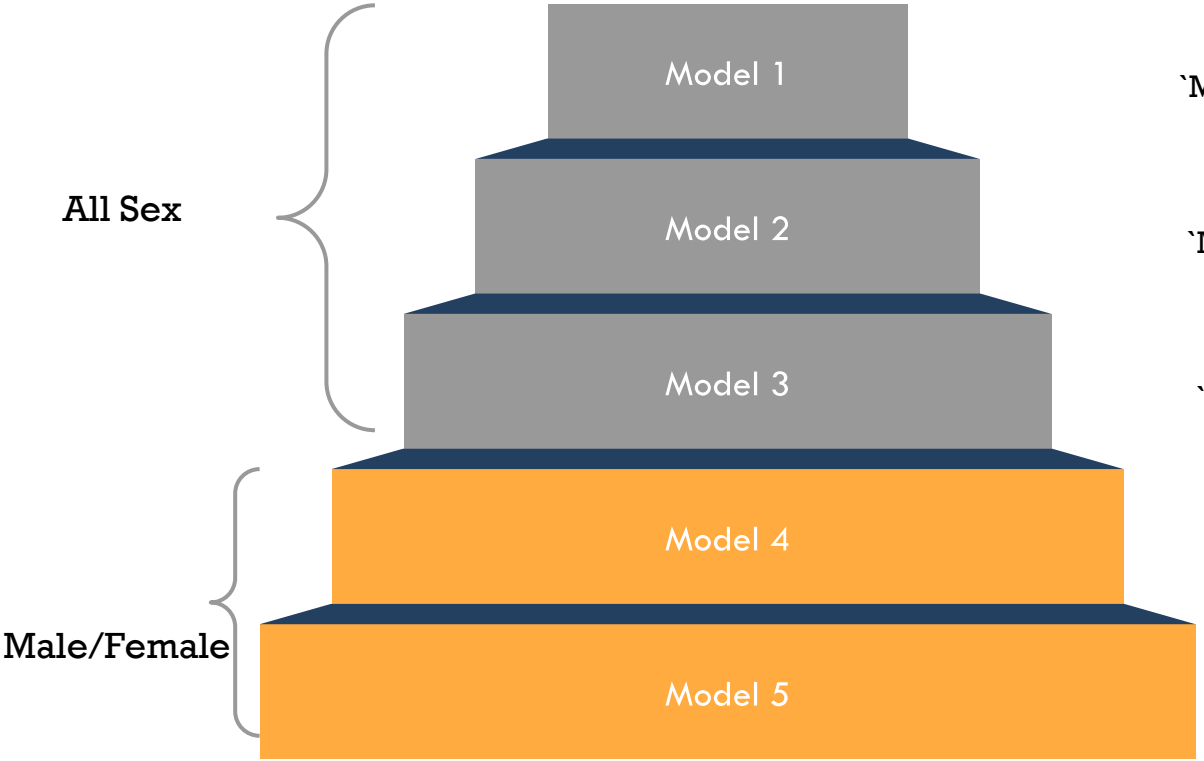


Female



Male

Modeling



`MRD-1 (L)` + `PTB (L)` + Age + Sex

`MRD-1 (L)` * `TPS (L)` * `PTB (L)` + Age + Sex

`MRD-1 (L)` * `TPS (L)` * `PTB (L)` * Age * Sex

`MRD-1 (L)` + `PTB (L)` + Age

Model 1: `MRD-1 (L)` * `TPS (L)` * `PTB (L)` + Age + Sex

# of Training observations: 143

# of Testing observations: 36

Sample Training data Summary Table

```
glm(formula = Ethnicity ~ `MRD-1 (L)` + `PTB (L)` + Age + Sex,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8925  -0.9555  -0.5393   1.0273   2.0754

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.65652    1.62835   1.631 0.102803
`MRD-1 (L)` -1.15581    0.29726  -3.888 0.000101 ***
`PTB (L)`    0.25014    0.07315   3.420 0.000627 ***
Age         -0.05239    0.01909  -2.744 0.006061 **
SexM         0.53622    0.52480   1.022 0.306893
---
```

CV 1

	FALSE	TRUE
FALSE	13	8
TRUE	6	9

CV 2

	FALSE	TRUE
FALSE	17	7
TRUE	1	11

CV 3

	FALSE	TRUE
FALSE	13	9
TRUE	6	8

CV 4

	FALSE	TRUE
FALSE	17	6
TRUE	2	11

Model 2: `MRD-1 (L)` + `PTB (L)` + Age + Sex

of Training observations: 143
# of Testing observations: 36

Sample Training data Summary Table

```
glm(formula = Ethnicity ~ `MRD-1 (L)` * `TPS (L)` * `PTB (L)` +
    Age + Sex, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8356  -1.0323  -0.3919   1.0112   2.3766

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.020116  15.159539  -0.133   0.8940
`MRD-1 (L)`                   -2.112333   4.434989  -0.476   0.6339
`TPS (L)`                     -1.413245   2.827919  -0.500   0.6173
`PTB (L)`                      0.898603   1.099691   0.817   0.4138
Age                           -0.032128   0.017878  -1.797   0.0723 .
SexM                           0.436470   0.534445   0.817   0.4141
`MRD-1 (L)`:`TPS (L)`          0.795761   0.842974   0.944   0.3452
`MRD-1 (L)`:`PTB (L)`         -0.040935   0.314959  -0.130   0.8966
`TPS (L)`:`PTB (L)`            0.007423   0.186018   0.040   0.9682
`MRD-1 (L)`:`TPS (L)`:`PTB (L)` -0.028620   0.054818  -0.522   0.6016
---
```

CV 1

	FALSE	TRUE
FALSE	16	7
TRUE	2	11

CV 2

	FALSE	TRUE
FALSE	14	7
TRUE	8	7

CV 3

	FALSE	TRUE
FALSE	14	5
TRUE	4	13

CV 4

	FALSE	TRUE
FALSE	11	2
TRUE	13	10

Model 3: `MRD-1 (L)` * `TPS (L)` * `PTB (L)` * Age * Sex

Sample Training data Summary Table

```
glm(formula = Ethnicity ~ `MRD-1 (L)` * `TPS (L)` * `PTB (L)` *
    Age * Sex, family = "binomial", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01522  -0.62528  -0.00001   0.44251   2.31912

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -1.983e+01  2.440e+02  -0.081    0.935
`MRD-1 (L)`                              1.846e+01  7.055e+01   0.262    0.794
`TPS (L)`                               -9.180e+00  4.481e+01  -0.205    0.838
`PTB (L)`                                4.446e+00  1.687e+01   0.264    0.792
Age                                      1.080e+00  4.630e+00   0.233    0.816
SexM                                     2.583e+02  9.622e+06   0.000    1.000
`MRD-1 (L)`:`TPS (L)`                   -1.797e-01  1.284e+01  -0.014    0.989
`MRD-1 (L)`:`PTB (L)`                   -2.181e+00  4.866e+00  -0.448    0.654
`TPS (L)`:`PTB (L)`                      4.748e-01  2.976e+00   0.160    0.873
`MRD-1 (L)`:Age                         -5.405e-01  1.344e+00  -0.402    0.688
`TPS (L)`:Age                           -1.562e-01  8.430e-01  -0.185    0.853
`PTB (L)`:Age                           -1.118e-01  3.130e-01  -0.357    0.721
`MRD-1 (L)`:SexM                        -1.764e+02  2.360e+06   0.000    1.000
`TPS (L)`:SexM                          -4.423e+02  2.517e+06   0.000    1.000
`PTB (L)`:SexM                          -2.755e+02  6.961e+05   0.000    1.000
Age:SexM                                -1.214e+02  2.169e+05  -0.001    1.000
`MRD-1 (L)`:`TPS (L)`:`PTB (L)`          7.667e-02  8.534e-01   0.090    0.928
`MRD-1 (L)`:`TPS (L)`:Age                8.324e-02  2.431e-01   0.342    0.732
`MRD-1 (L)`:`PTB (L)`:Age                4.884e-02  9.080e-02   0.538    0.591
`TPS (L)`:`PTB (L)`:Age                  1.010e-02  5.439e-02   0.186    0.853
`MRD-1 (L)`:`TPS (L)`:SexM               1.777e+02  6.343e+05   0.000    1.000
`MRD-1 (L)`:`PTB (L)`:SexM               7.965e+01  1.666e+05   0.000    1.000
`TPS (L)`:`PTB (L)`:SexM                 1.009e+02  1.858e+05   0.001    1.000
`MRD-1 (L)`:Age:SexM                     3.521e+01  5.307e+04   0.001    0.999
`TPS (L)`:Age:SexM                       3.887e+01  5.494e+04   0.001    0.999
`PTB (L)`:Age:SexM                       1.678e+01  1.578e+04   0.001    0.999
`MRD-1 (L)`:`TPS (L)`:`PTB (L)`:Age     -6.027e-03  1.575e-02  -0.383    0.702
`MRD-1 (L)`:`TPS (L)`:`PTB (L)`:SexM    -3.180e+01  4.595e+04  -0.001    0.999
`MRD-1 (L)`:`TPS (L)`:Age:SexM          -1.186e+01  1.374e+04  -0.001    0.999
`MRD-1 (L)`:`PTB (L)`:Age:SexM          -4.741e+00  3.767e+03  -0.001    0.999
`TPS (L)`:`PTB (L)`:Age:SexM            -4.948e+00  4.070e+03  -0.001    0.999
`MRD-1 (L)`:`TPS (L)`:`PTB (L)`:Age:SexM  1.465e+00  9.987e+02   0.001    0.999
```

# of Training observations: 143

# of Testing observations: 36

CV 1

	FALSE	TRUE
FALSE	12	11
TRUE	7	6

CV 2

	FALSE	TRUE
FALSE	13	6
TRUE	6	11

CV 3

	FALSE	TRUE
FALSE	16	4
TRUE	6	10

CV 4

	FALSE	TRUE
FALSE	14	6
TRUE	4	12

Model 4: `MRD-1 (L)` + `PTB (L)` + Age (Female)

of Training observations: 116

# of Testing observations: 29

Sample Training data Summary Table

```
glm(formula = Ethnicity ~ `MRD-1 (L)` + `PTB (L)` + Age, family = "binomial",
    data = trainFemale)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8788  -0.9213  -0.4312   0.9600   2.0432

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.45103    1.83641   2.424 0.015360 *
`MRD-1 (L)` -1.30805    0.33767  -3.874 0.000107 ***
`PTB (L)`    0.22464    0.07957   2.823 0.004756 **
Age         -0.06759    0.02193  -3.082 0.002054 **
---
```

CV 1

	FALSE	TRUE
FALSE	13	4
TRUE	4	8

CV 2

	FALSE	TRUE
FALSE	14	4
TRUE	2	9

CV 3

	FALSE	TRUE
FALSE	11	4
TRUE	4	10

CV 4

	FALSE	TRUE
FALSE	12	6
TRUE	4	7

Model 5: `MRD-1 (L)` + `PTB (L)` + Age (Male)

of Training observations: 27

# of Testing observations: 7

Sample Training data Summary Table

```
glm(formula = Ethnicity ~ `MRD-1 (L)` + `PTB (L)` + Age, family = "binomial",
    data = trainMale)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5348  -0.8073  -0.4046   0.9015   1.9734

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.64960    3.45440  -1.057    0.291
`MRD-1 (L)` -0.91818    0.64764  -1.418    0.156
`PTB (L)`    0.45805    0.22187   2.064    0.039 *
Age          0.01087    0.04227   0.257    0.797
```

CV 1

	FALSE	TRUE
FALSE	2	1
TRUE	2	2

CV 2

	FALSE	TRUE
FALSE	3	2
TRUE	0	2

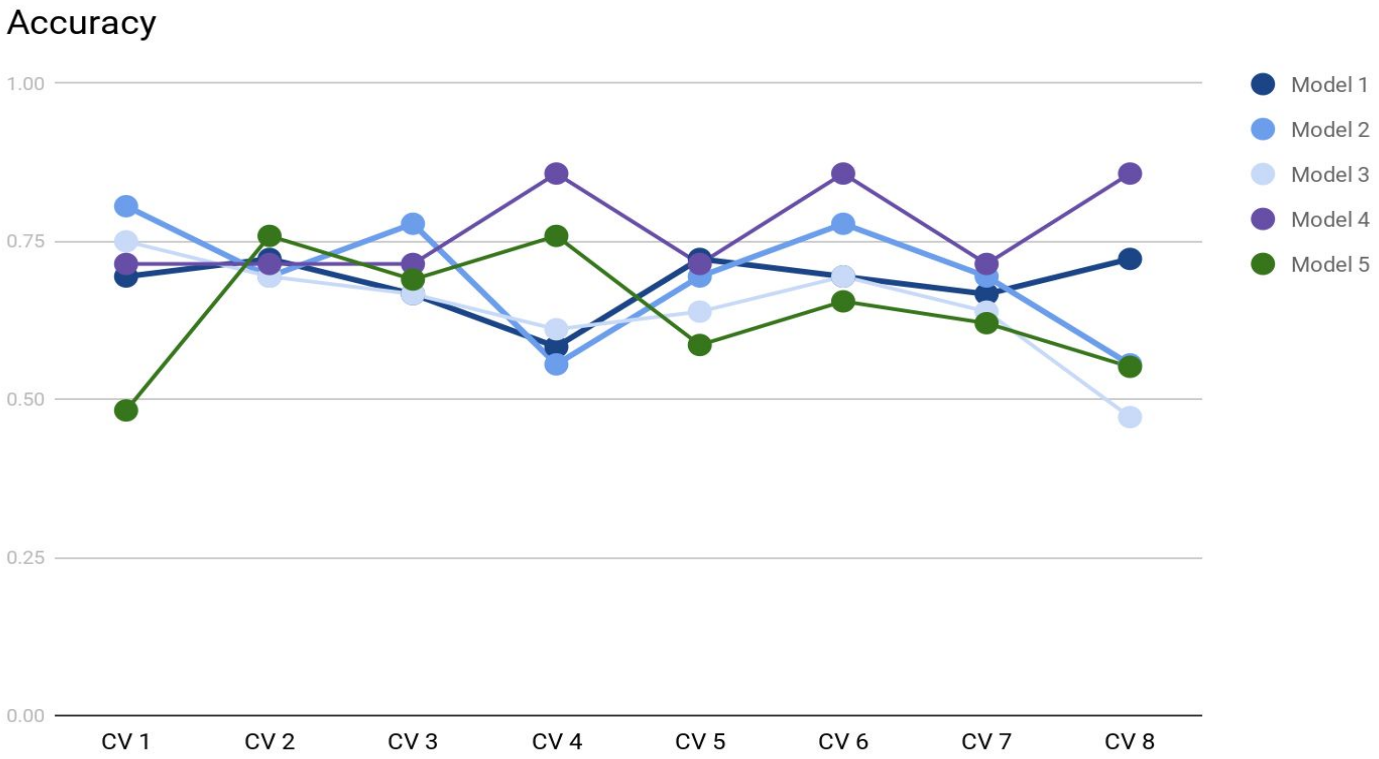
CV 3

	FALSE	TRUE
FALSE	4	2
TRUE	0	1

CV 4

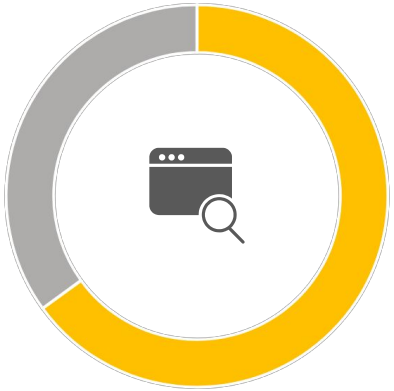
	FALSE	TRUE
FALSE	5	0
TRUE	1	1

Result



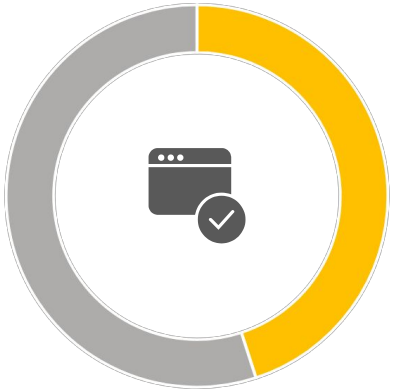
- high uncertainty due to small data size
- Add validation dataset

Artificial Neural Network



ANN Performance

Training Accuracy: 0.65
Validation Accuracy: 0.70



Compare to other Models

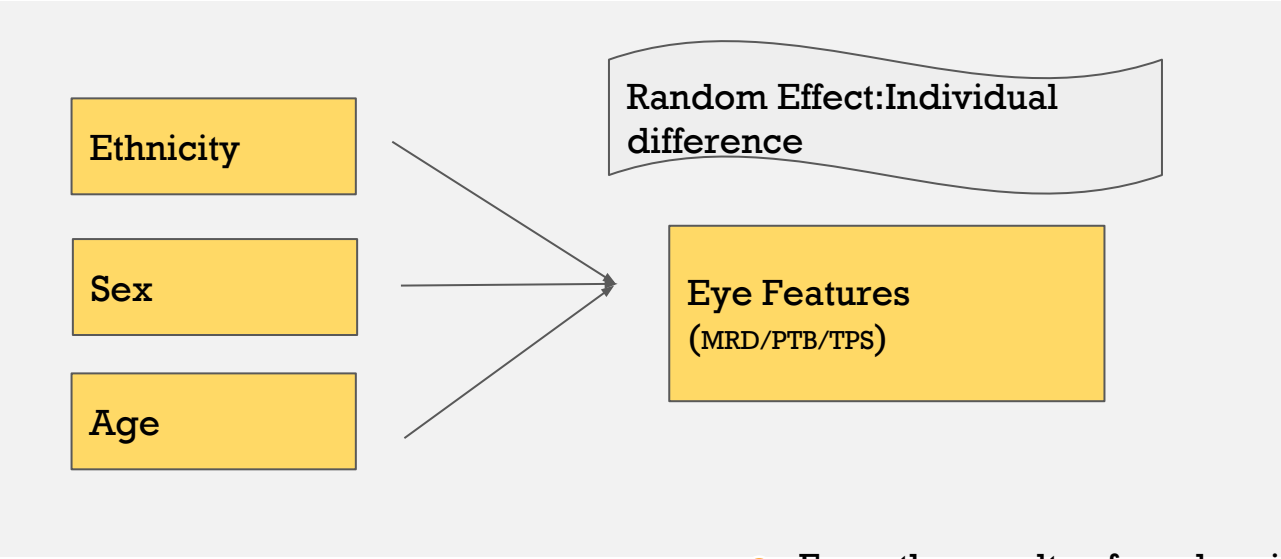
ANN model has similar performance as our simpler models

ANN part

To serve as a comparison to the logistic regression model, we decided to sacrifice all interpretability and run an artificial neural network model to learn a mapping from age, sex, and all metrics of eyes to the ethnicity of the patient.
In order to do this, we use an ANN with 1 input layer, 4 hidden layer, and 1 output layer. For each of the hidden layer, we choose to use 64, 128, 32, and 16 dimensions. Overall, this would give us a complex model with thousands of trainable parameters. Although the model learned from the dataset, we have is not interpretable (since the ANN is like a black box, and there are just too many parameters to track), ANN usually provides us with the ultimate flexibility and accuracy. Any derived parameters that would work can be automatically learned by a neural network.

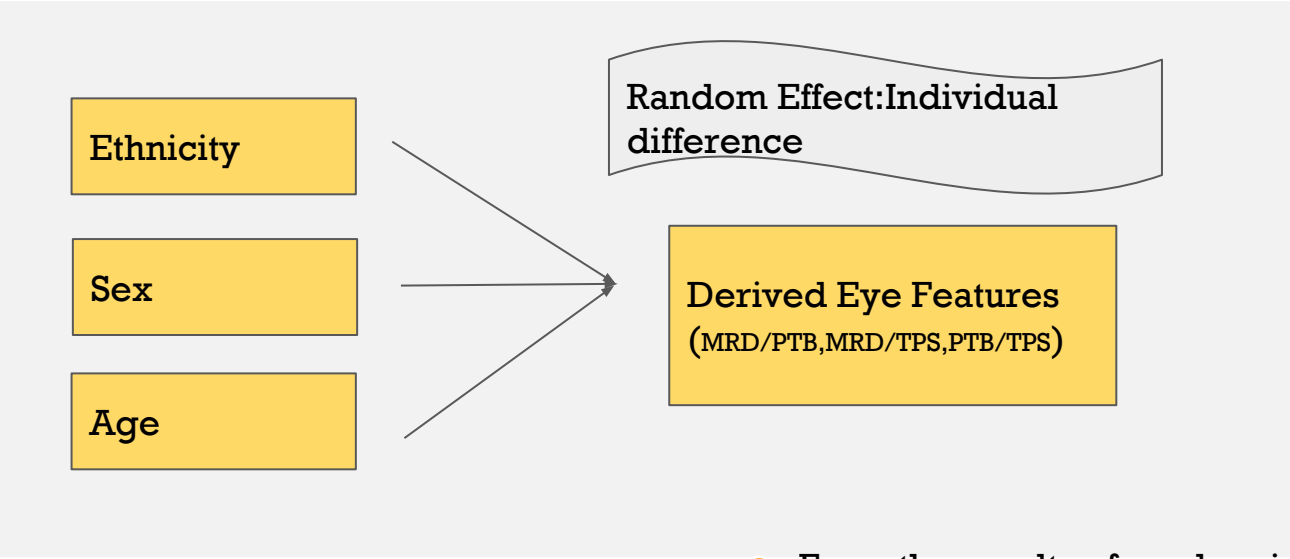
In terms of the result we are getting, it, however, turns out that the performance of ANN is similar to our simple logistic models. This indicates that there are significant noises in the dataset, and the best we could do in predicting the ethnicity is about 70%. Moreover, the fact that the simple logistic model actually performs similarly or even better than more complex models such as ANN indicates that the true relationship between ethnicity and metrics of the eyes should be simple, and an extremely naïve model is likely the solution to the problem.

Can being Persian Affect the eye features?



- From the results of random intercept model, the p-value of three predictors are significant
- Sex and Age plays an important role to decide eye features

Can being Persian Affect the eye features?



- From the results of random intercept model, the Ethnicity is significant for MRD/PTB and MRD/TPS.
- Sex and Age plays an important role to decide eye features

Random Effect Model Part

In this part of project, we are solving the question "how the ethnicity of a patient affects the metrics of their eyes". In this part, we explore how all of the three metrics (MRD, PTB, TPS) and three derived metrics (MRD/PTB, MRD/TPB, PTB/TPS). Based on our result, we found that in fact that the ethnicity is a significant predictor for metrics MRD, PTB, TPS, MRD/PTB, and MRD/TPB.

<u>Fitted models are the following</u>:

- Notice that effect of Age is small, since it is expected that patients who are just 1 year older are not going to have big difference in their eyes

MRD = 4.8021 - 0.3448 * **Ethnicity is Persian** - 0.2629 * **Sex is Male** - 0.0210 * **Age**

PTB = 12.4018 - 1.4596 * **Ethnicity is Persian** - 3.0753 * **Sex is Male** - 0.0432 * **Age**

TPS = 3.7385 - 0.7382 * **Ethnicity is Persian** - 1.0937 * **Sex is Male** - 0.0174 * **Age**

MRD/TPS = 1.4241 - 0.2900 * **Ethnicity is Persian** + 0.2358 * **Sex is Male** - 0.0091 * **Age**

MRD/PTB = 0.3773 - 0.0549 * **Ethnicity is Persian** + 0.0420 * **Sex is Male** - 0.0022 * **Age**

PTB/TPS = 3.6112 - 0.2544 * **Ethnicity is Persian** + 0.1317 * **Sex is Male** - 0.0017 * **Age**

We summarize the interpretation (of Ethnicity) we draw from our models in this form:

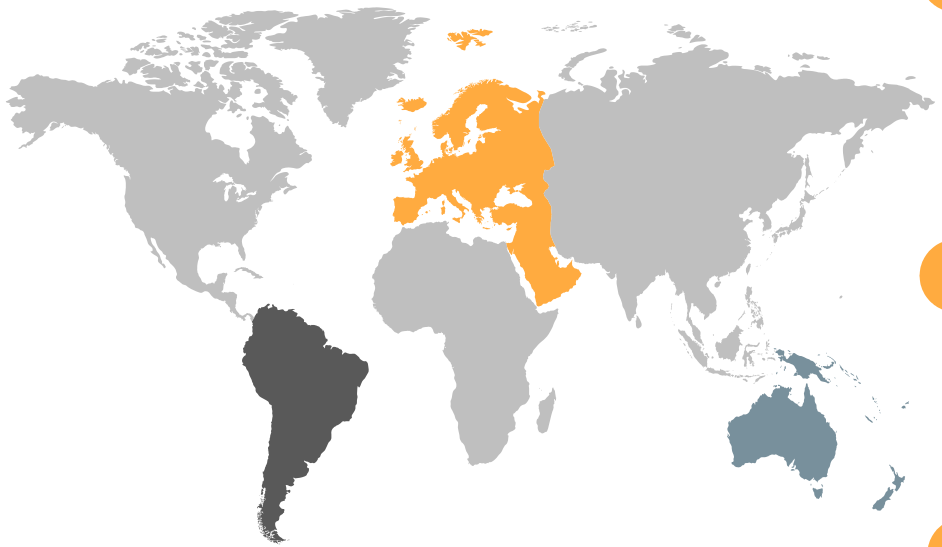
Metrics	Ethnicity's Coefficient	P-value	Significant	Interpretation
MRD	-0.3448	0.0005	Yes	At a particular age and with a particular gender, a patient who is Persian is expected to have an MRD that is 0.3448 unit lower than patient who is Caucasian
PTB	1.4596	0.0004	Yes	At a particular age and with a particular gender, a patient who is Persian is expected to have a PTB that is 1.4596 unit higher than patient who is Caucasian
TPS	0.7382	0.0046	Yes	At a particular age and with a particular gender, a patient who is Persian is expected to have a TPS that is 0.7382 unit higher than patient who is Caucasian
MRD/TPS	-0.2900	0.0000	Yes	At a particular age and with a particular gender, a patient who is Persian is expected to have an MRD/TPS ratio that is 0.29 lower than patient who is Caucasian
MRD/PTB	-0.0549	0.0000	Yes	At a particular age and with a particular gender, a patient who is Persian is expected to have an MRD/PTB ratio that is 0.0549 lower than patient who is Caucasian
PTB/TPS	-0.2544	0.1427	No	At a particular age and with a particular gender, a patient who is Persian is expected to have a PTB/TPS ratio that is 0.2544 lower than patient who is Caucasian

We notice that from the result of this part, it seems contradictory with respect to our logistic modeling sections, as a lot of the metrics are not significant in the logistic regression model. The reasons are the following:

1. In logistic models, the parameters can be affected by multicollinearity, which would inflate the variances of coefficient and cause some of them to appear as insignificant.

2. In logistic models, we did not exclude the noise (i.e. the random effect of patient). Thus, the variances of the coefficients will be a combination of variances due to random effect as well as variances due to the predictors themselves. This cause some of the predictors to become insignificant. Notice, we did not exclude the random effect in our logistic regression model part since the question we are attempting to answer require a predictive model. Thus, we do not seek to exclude those random effect as they won't impact the coefficients in the model and won't make any improvement to model accuracies.

Conclusion



Both ANN and logistic models failed to achieve a great validation score in predicting the Ethnicity, indicating there could be high noises.

Significant split observed with random intercept model (patient as random effect) indicates Ethnicity makes a differences in many metrics

Although ethnicity will split various metrics of a eye, this differences in the metrics will vary with each patient


THANK YOU