Enova Technical Assessment Solution

Data Loading & Evaluation

Use Python pandas and numpy packages to load data and do some basic manipulation.

Check the dataset statistics summary and find out anything interesting.

- There are some missing values(NaN)
- Feature 'symptoms' is a tricky feature, as each one has multiple values
- Sanity check, is the data balanced? Eg: if survival_1_years = 0, then survival_7_years must be 0
- How are the features distributed?
- Are the features highly correlated?
- Which features are important?

EDA (Exploratory Data Analysis)

This section I solved the questions from last section.

- For missing values I set it into 0 or mean value depends on the feature itself
- For 'symptoms', I split the nultiple values first, then do one hot encoding
- The labels(survival_7_years) is almost balanced with ratio 0.43. And the survival_1_years, survival_7_years don't conflict with each other
- The training data and score data's features followed the same distribution. So the training data should reflect the score data set condition
- There are some features correlated with each other, but not a very big problem here
- After running machine learning model, I gave the feature importance plot as well which showed the symptoms features are the most important ones

Model Selection

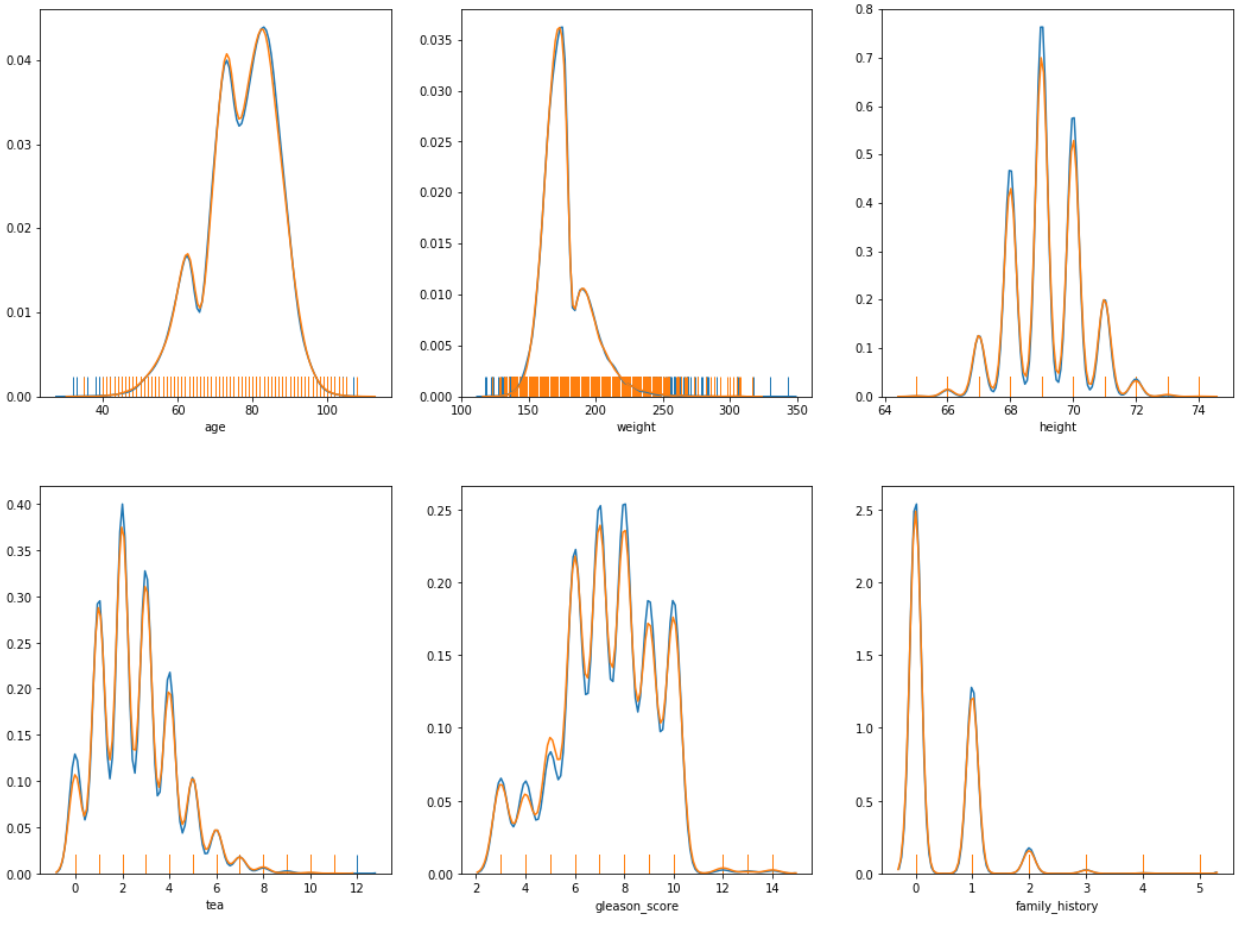
There are many classification models to choose, we need to know that the dataset has a lot of categorical features. So we don't consider logistics regression which is not good dealing with too many categorical data.

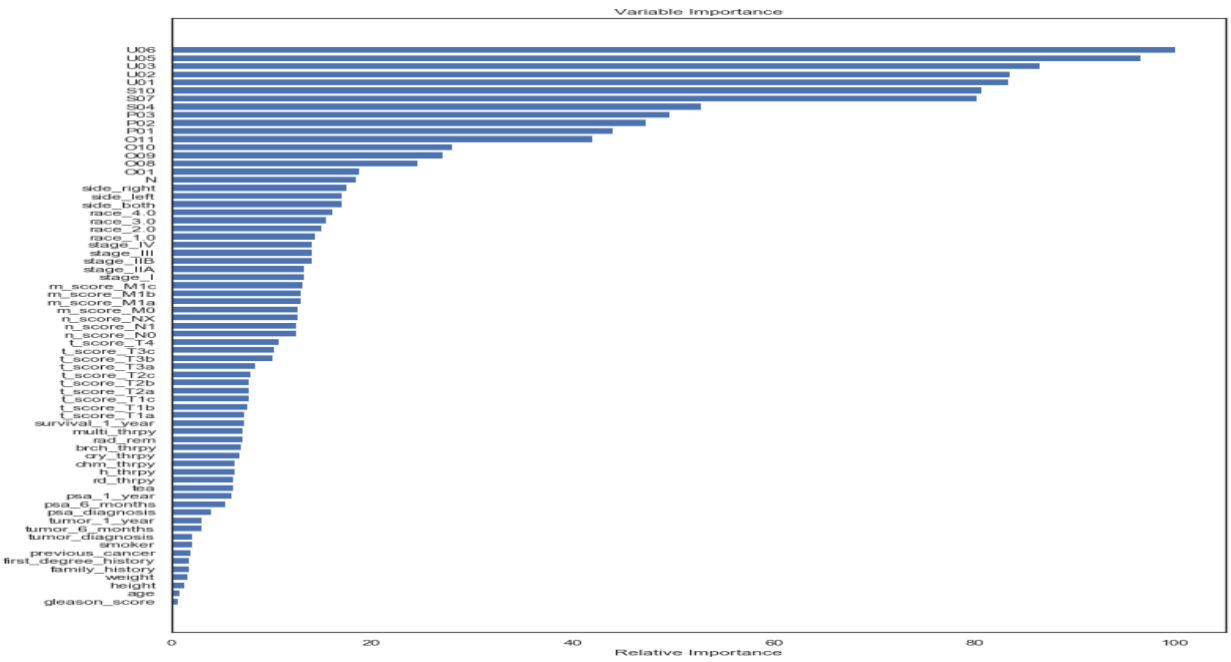
Instead, I picked random forest(ensemble method), xgboost(ensemble method), svm, svm-smote to train the model respectively.

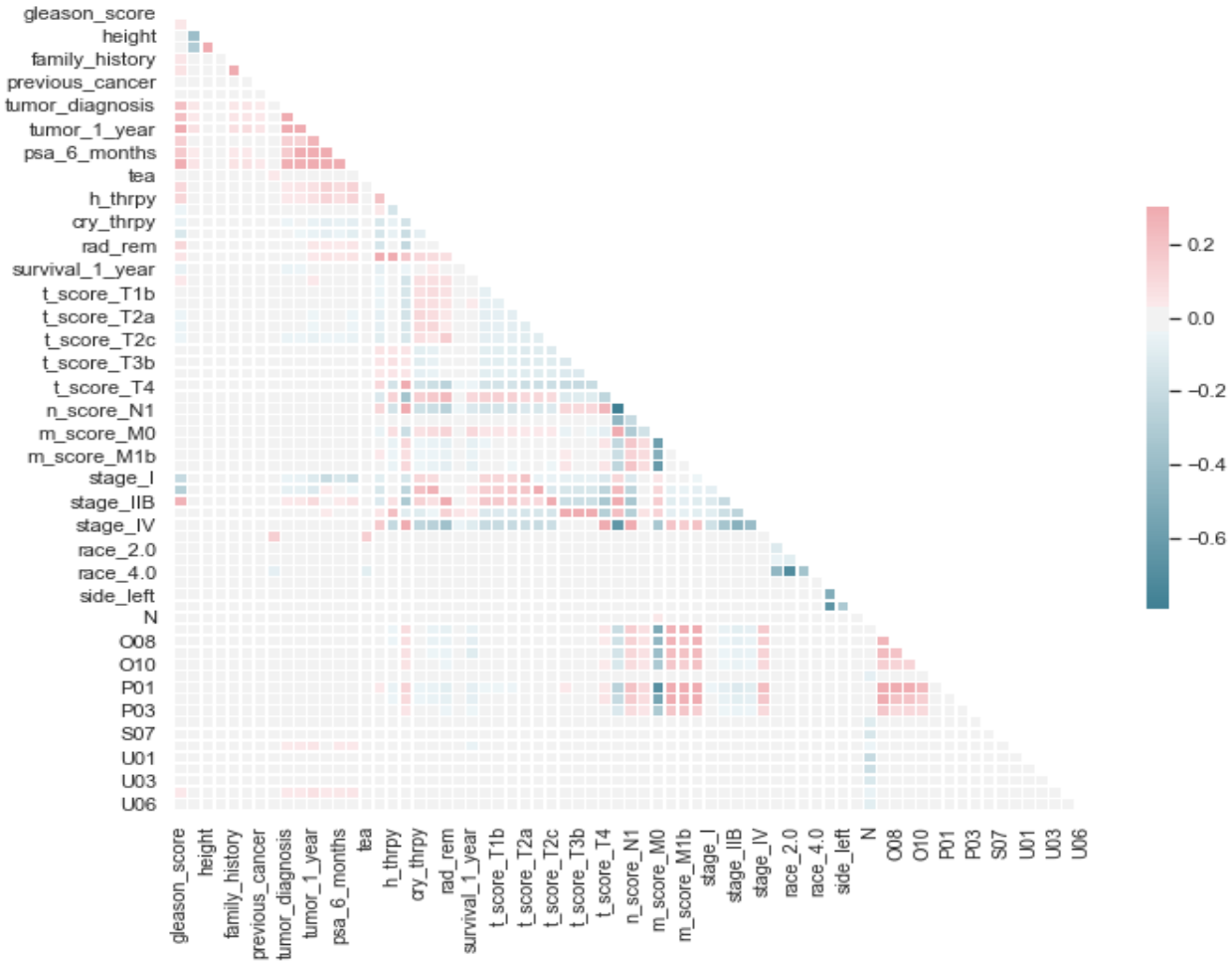
Random forest is an ensemble method which can reduce the model bias and unlikely to overfit. Xgboost is a boosting model which are quite fast and robust. Svm is good for model which has a lot of features. As the data is not super balanced, I tried svm-smote to automatically fixed the imbalance data issue.

Finally I picked Xgboost as final model which has relatively highest performance which has test set accuracy 68.35%.

Some Data Visualizations




Variable Importance
U06
U05
U03
U02
U01
S10
S07
S04
P03
P02
P01
O11
O10
O09
O08
O01
N
side_right
side_left
side_both
race_4.0
race_3.0
race_2.0
race_1.0
stage_IV
stage_III
stage_IIB
stage_IIA
stage_I
m_score_M1c
m_score_M1b
m_score_M1a
m_score_M0
n_score_NX
n_score_N1
n_score_N0
t_score_T4
t_score_T3c
t_score_T3b
t_score_T3a
t_score_T2c
t_score_T2b
t_score_T2a
t_score_T1c
t_score_T1b
t_score_T1a
survival_1_year
multi_thrpy
rad_rem
brch_thrpy
cry_thrpy
chm_thrpy
h_thrpy
rd_thrpy
tea
psa_1_year
psa_6_months
psa_diagnosis
tumor_1_year
tumor_6_months
tumor_diagnosis
smoker
previous_cancer
first_degree_history
family_history
weight
height
age
gleason_score
0
20
40
60
80
100
Relative Importance


gleason_score
height
family_history
previous_cancer
tumor_diagnosis
tumor_1_year
psa_6_months
tea
h_thrpy
cry_thrpy
rad_rem
survival_1_year
t_score_T1b
t_score_T2a
t_score_T2c
t_score_T3b
t_score_T4
n_score_N1
m_score_M0
m_score_M1b
stage_I
stage_IIB
stage_IV
race_2.0
race_4.0
side_left
N
O08
O10
P01
P03
S07
U01
U03
U06
0.2
0.0
−0.2
−0.4
−0.6