CHRONIC KIDNEY DISEASE

Elliot Lee

TABLE OF CONTENTS

01 Introduction

02 Data Exploration

03 Feature Engineering & Modelling

04 Conclusion

01

Introduction

INTRODUCTION

Chronic Kidney Disease (CKD) is the progressive loss of kidney function over a period of several years

- May lead to permanent kidney failure
- 5 stages
- Diabetes and high blood pressure are leading causes

PROBLEM STATEMENT

Classification Model

Predict if a patient will progres in CKD staging given longitudinal lab measurements

Metrics

ROC-AUC and Recall

Key target: Identifying the positive class

Targeted Intervention

Allows for earlier identification of patients who may progress in staging and hence earlier intervention

02

Data Exploration

Datasets

Number of records
300

Missing data
No

Datasets
9

Predictors
Lab measurements, demographics, drug intake history

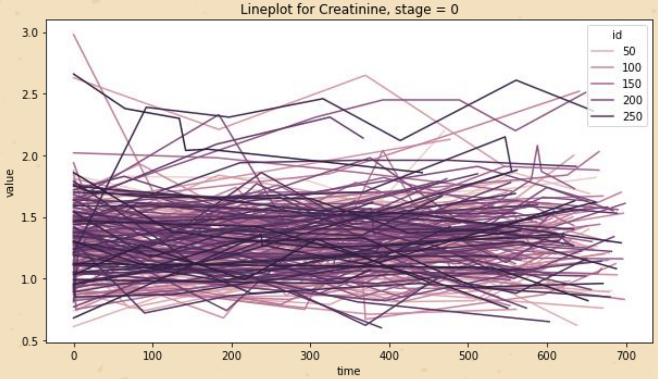
Target variable

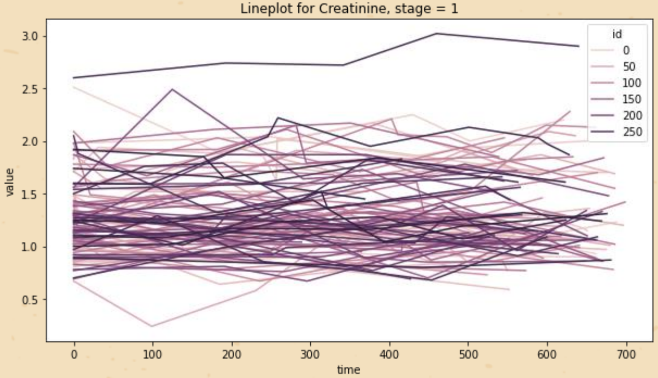
1 – will progress in CKD
0 – will not progress in CKD

Target variable distribution

1 – 100 (33.3%),
0 – 200 (66.7%)

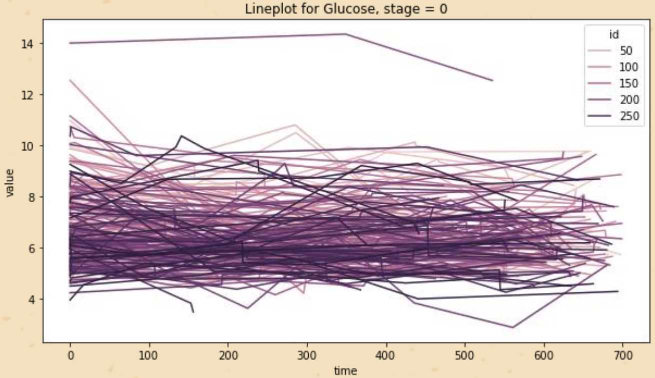
Data Exploration

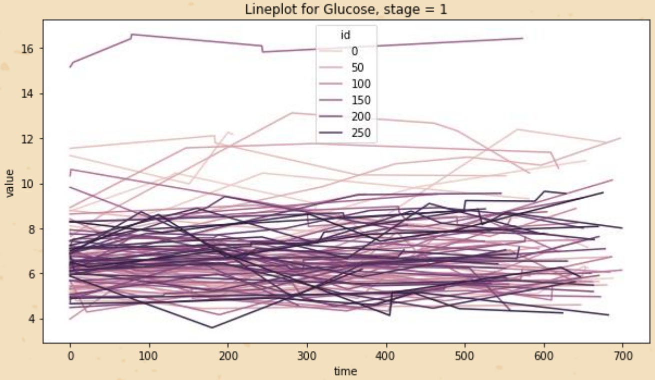




Data is largely stationary for health parameters

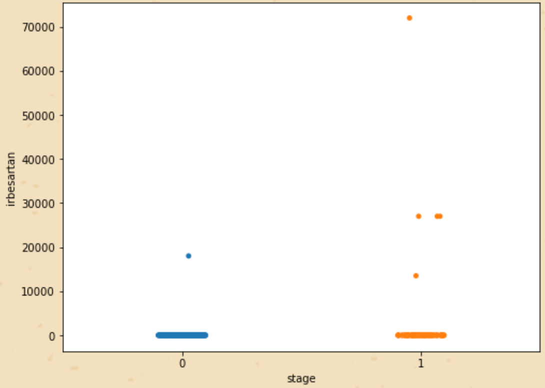
Data Exploration

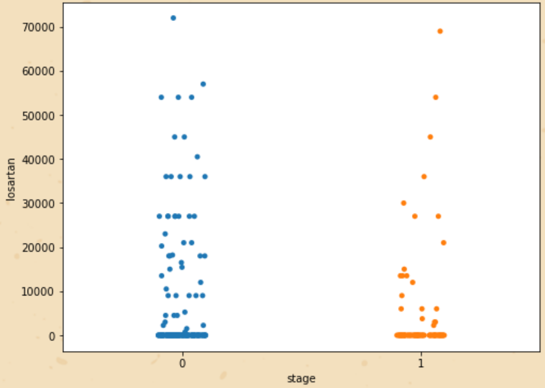




Some outlier values for glucose

Data Exploration





Many of the drugs did not have a signifcant number of users

Some were more prevalent amongst patients, but did not show observable differences

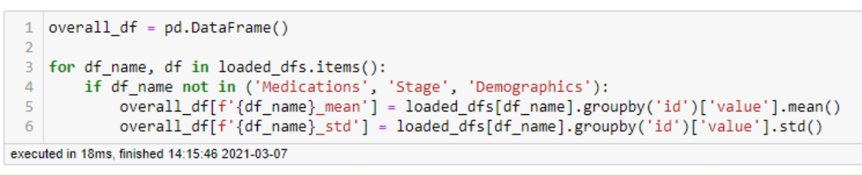
03

Feature Engineering & Modelling

Feature Engineering

Numeric Parameters

Condensed into mean value
Obtained standard deviation as a measure of the fluctuations

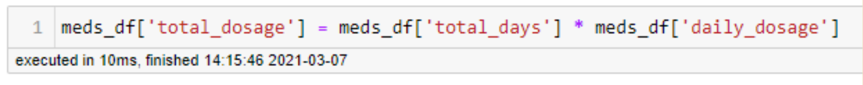


```
overall_df = pd.DataFrame()

for df_name, df in loaded_dfs.items():
    if df_name not in ('Medications', 'Stage', 'Demographics'):
        overall_df[f'{df_name}_mean'] = loaded_dfs[df_name].groupby('id')['value'].mean()
        overall_df[f'{df_name}_std'] = loaded_dfs[df_name].groupby('id')['value'].std()
```

```
executed in 18ms, finished 14:15:46 2021-03-07
```

Drugs

Calculated overall dosage



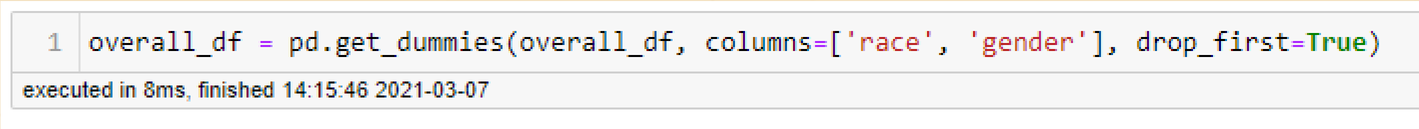
```
meds_df['total_dosage'] = meds_df['total_days'] * meds_df['daily_dosage']
```

```
executed in 10ms, finished 14:15:46 2021-03-07
```

Feature Engineering

Demographic Data

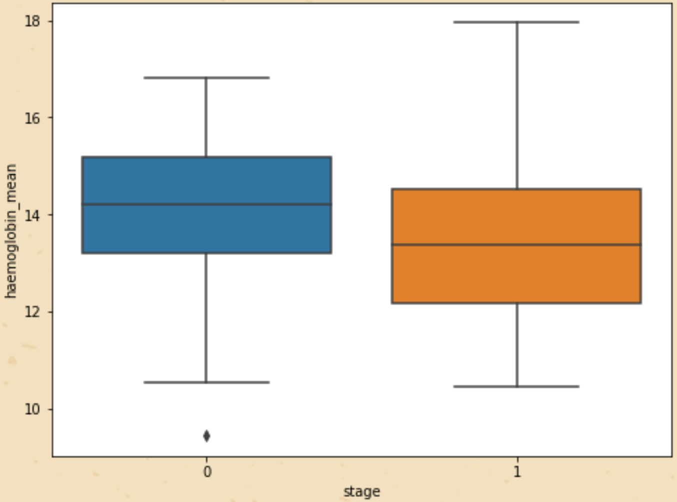
One-hot encoded



```
overall_df = pd.get_dummies(overall_df, columns=['race', 'gender'], drop_first=True)
```

```
executed in 8ms, finished 14:15:46 2021-03-07
```

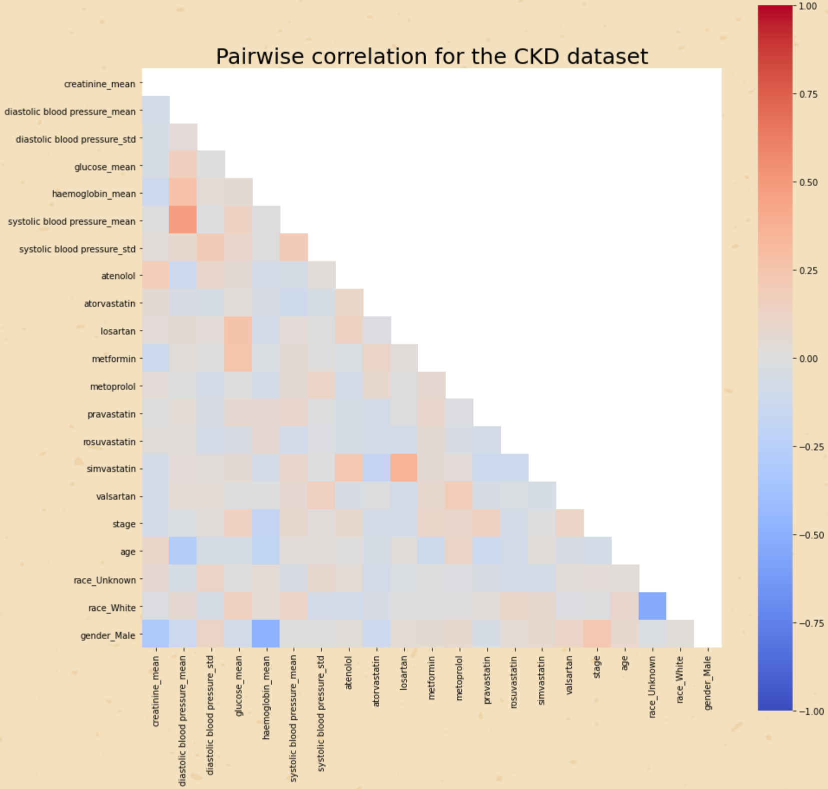
Feature Selection



Boxplot

Mean haemoglobin level of patients who progress in CKD staging is lower than those who do not

Feature Selection



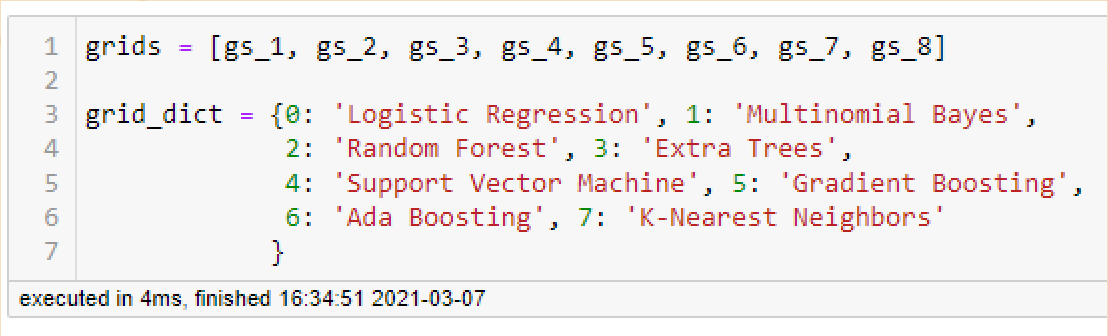
Correlation plot

Variables are largely uncorrelated except for the blood pressure measurements and race variables

Modelling

Classification Models

7 different classification models were chosen and GridSearchCV was used to obtain the best cross-validated recall score

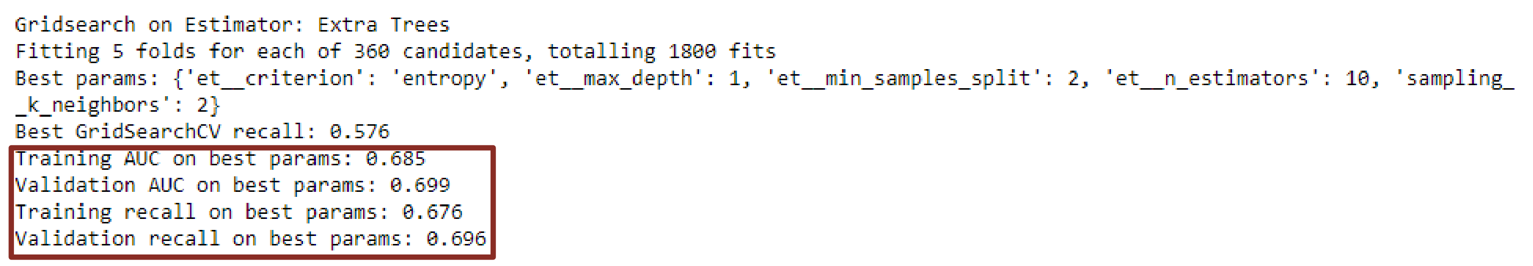


```
grids = [gs_1, gs_2, gs_3, gs_4, gs_5, gs_6, gs_7, gs_8]

grid_dict = {0: 'Logistic Regression', 1: 'Multinomial Bayes',
             2: 'Random Forest', 3: 'Extra Trees',
             4: 'Support Vector Machine', 5: 'Gradient Boosting',
             6: 'Ada Boosting', 7: 'K-Nearest Neighbors'
            }
```

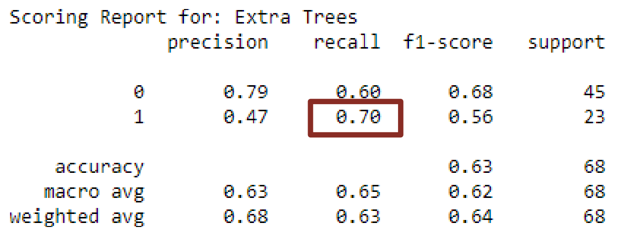
```
executed in 4ms, finished 16:34:51 2021-03-07
```

Modelling

Extra Trees Model



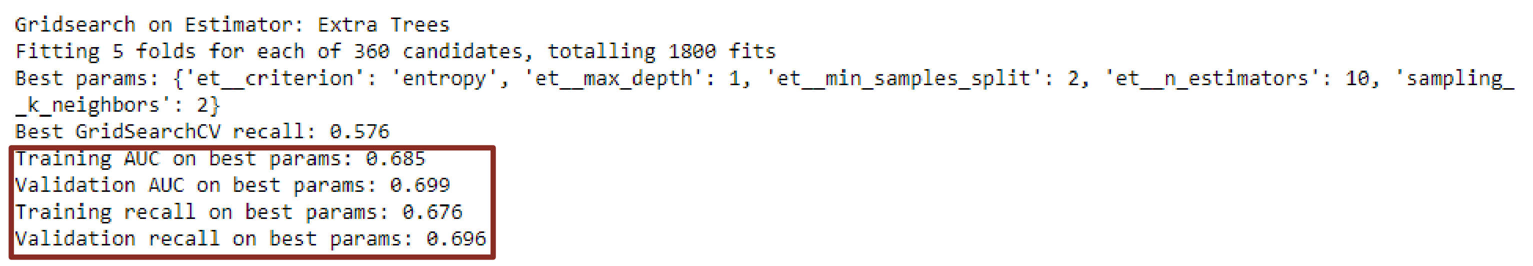
```
Gridsearch on Estimator: Extra Trees
Fitting 5 folds for each of 360 candidates, totalling 1800 fits
Best params: {'et__criterion': 'entropy', 'et__max_depth': 1, 'et__min_samples_split': 2, 'et__n_estimators': 10, 'sampling_
_k_neighbors': 2}
Best GridSearchCV recall: 0.576
Training AUC on best params: 0.685
Validation AUC on best params: 0.699
Training recall on best params: 0.676
Validation recall on best params: 0.696
```



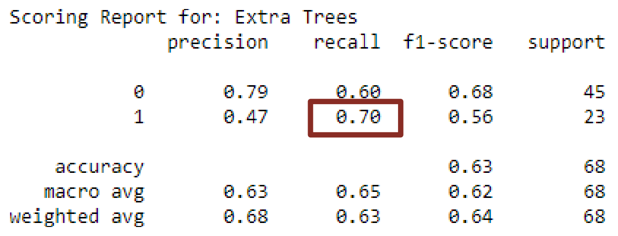
```
Scoring Report for: Extra Trees
              precision    recall  f1-score   support

           0       0.79      0.60      0.68        45
           1       0.47      0.70      0.56        23

    accuracy                           0.63        68
   macro avg       0.63      0.65      0.62        68
weighted avg       0.68      0.63      0.64        68
```

Modelling

Extra Trees Model

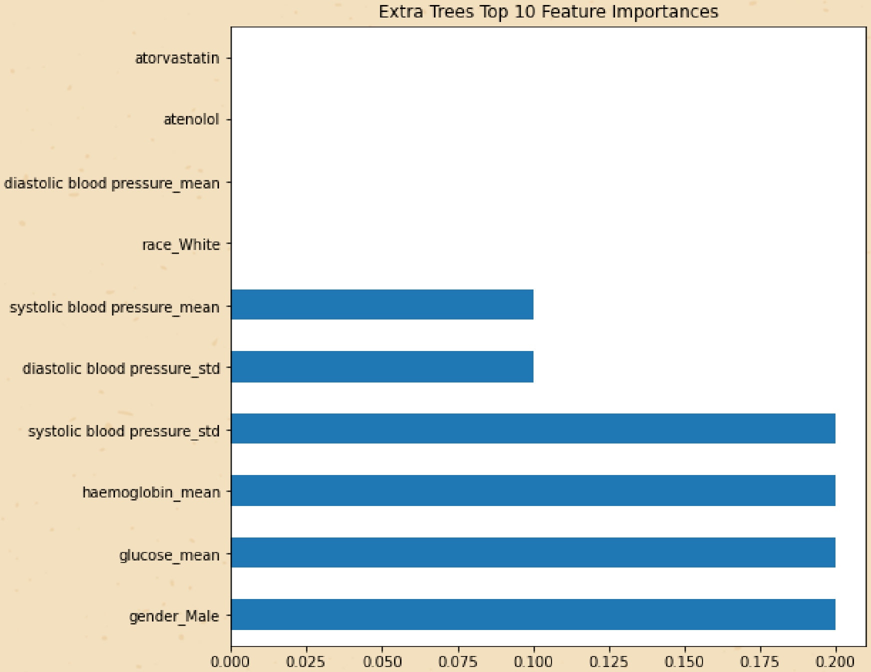


```
Gridsearch on Estimator: Extra Trees
Fitting 5 folds for each of 360 candidates, totalling 1800 fits
Best params: {'et__criterion': 'entropy', 'et__max_depth': 1, 'et__min_samples_split': 2, 'et__n_estimators': 10, 'sampling_
_k_neighbors': 2}
Best GridSearchCV recall: 0.576
Training AUC on best params: 0.685
Validation AUC on best params: 0.699
Training recall on best params: 0.676
Validation recall on best params: 0.696
```

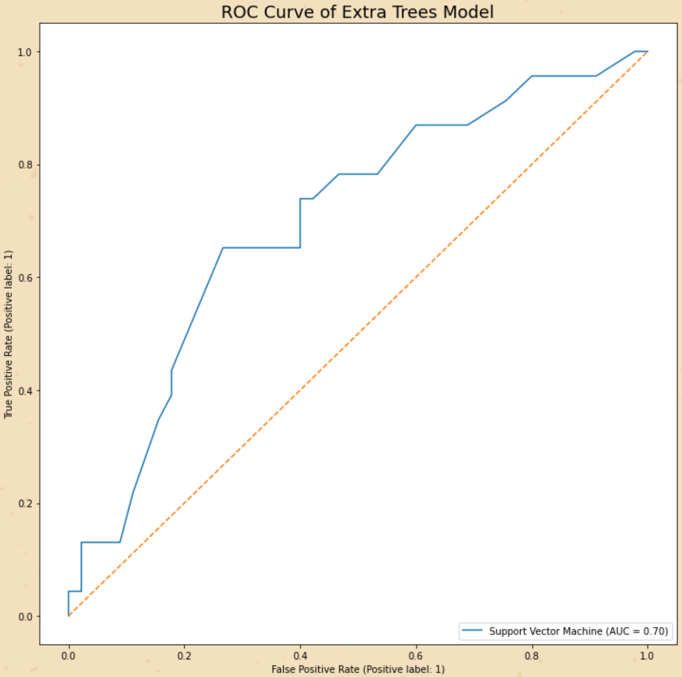


```
Scoring Report for: Extra Trees
              precision    recall  f1-score   support

           0       0.79      0.60      0.68        45
           1       0.47      0.70      0.56        23

    accuracy                           0.63        68
   macro avg       0.63      0.65      0.62        68
weighted avg       0.68      0.63      0.64        68
```

Model Evaluation



Extra Trees Top 10 Feature Importances



ROC Curve of Extra Trees Model

04

Conclusion

Conclusion

Classification Model

ROC-AUC of 0.699
Recall of 0.696

Targeted Intervention

Allows for better targeted intervention and treatment to halt or control the disease

Limitations

Recall and AUC are not extremely high

Deploy a mixed-effect model to better account for variability in efficacy of drugs

THANKS!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

Please keep this slide for attribution