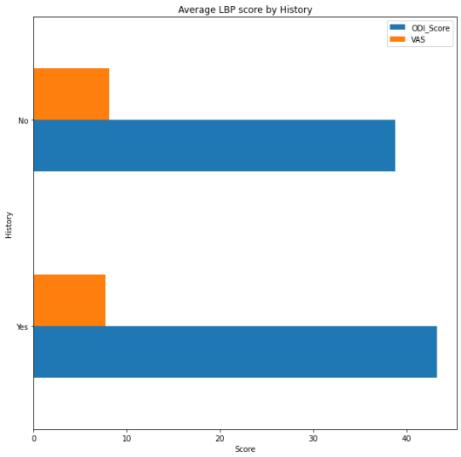
Capstone Report
Predicting Low Back Pain Severity

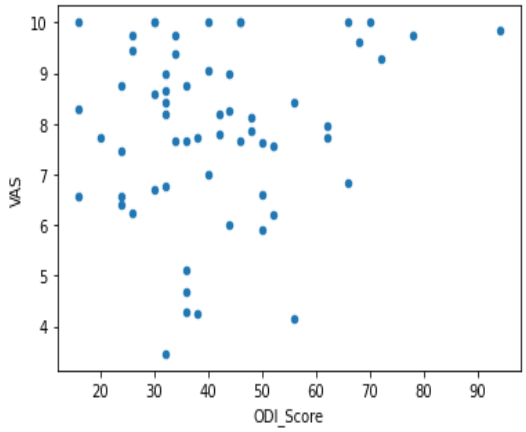
Introduction

LBP is one of the most common disorders affecting people. Previous research has identified three primary risk factor categories for LBP (occupational, personal and psychosocial factors). Using these factors to build a predictive model can help determine the severity and develop effective intervention methods.

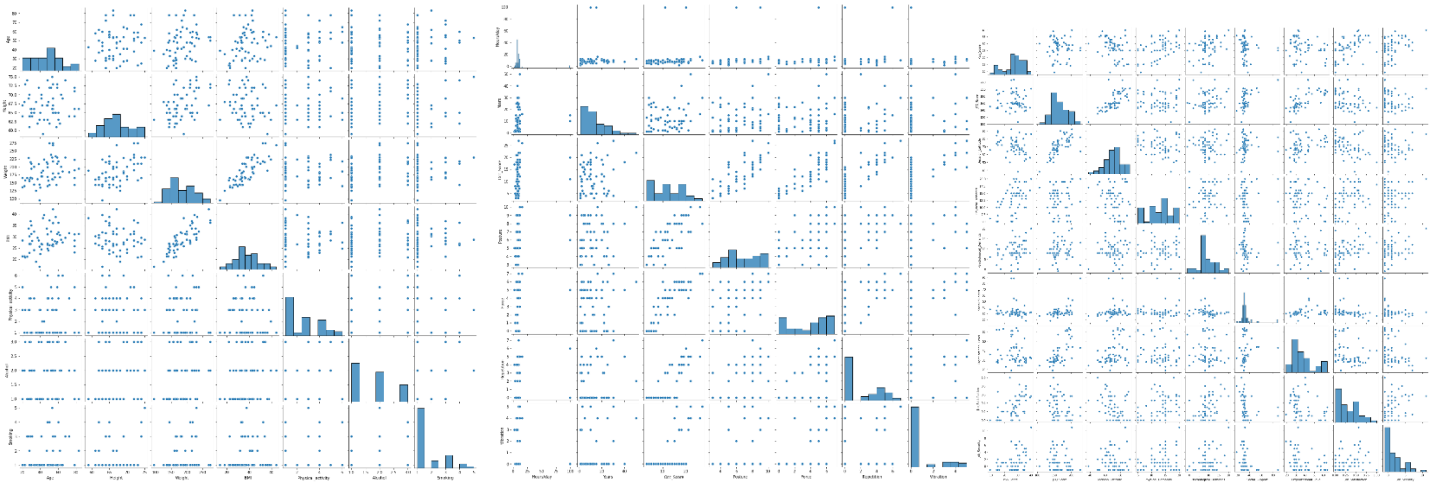
Exploratory Data Analysis

The data was collected as part of a dissertation study from 60 participants who reported some degree of low back pain. It consists of attributes from each of the three risk factor categories and 2 measures of LBP severity – VAS and ODI Score. Factors from all three categories were used to build the predictive model. The data contains 60 rows and 28 columns and contains information such as age, weight, height, BMI, physical activity, family history of LBP. Occupations factors such as force, repetition, posture, vibration etc. And psychosocial factors such as the Perceived Stress Score (PSS). Looking at the means, only the ODI Score was chosen as the dependent variable since VAS did not seem to be different between different groups observed.





Correlations were observed between some of the features in each risk category.



Feature Selection

Missing values were filled with median values. In order to apply a classification model, several features were converted into categorical variables. A binary and tertiary model was built for which 2 ODI features were created: ODI – split into 0,

1 and 2 for low, moderate and high LBP severity and ODI2 – split into 0 and 1 for low and high LBP severity. One hot encoding was performed on the categorical variables - Gender, History, Ethnicity, BMI, Physical activity etc.

Preprocessing

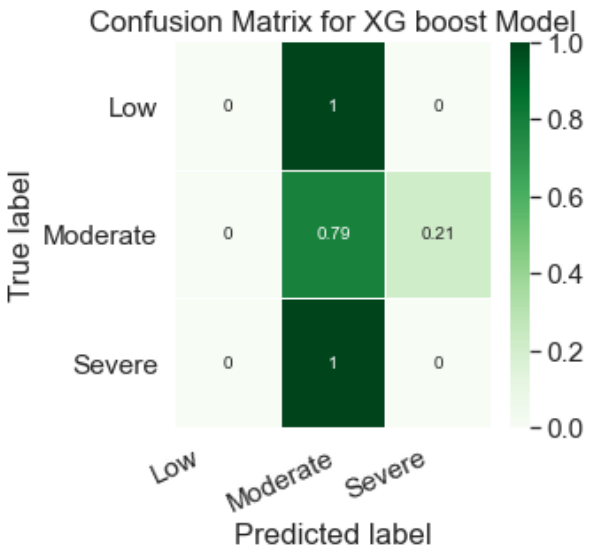
Information Value or IV was used in feature selection in the binary model. Those features thar exhibited statistics in the range of 0.1 to 0.8 were selected for model building. Further, a technique called Variance Inflation Factor (VIF) was used to reduce multicollinearity between the features. A VIF of greater than 5 indicates extreme multicollinearity and is avoided.

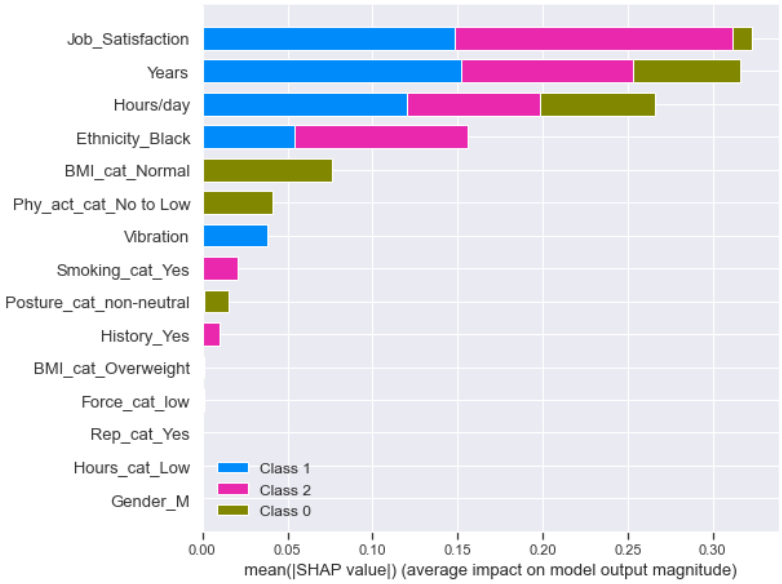
Modelling

The data is split into 70/30 for training and testing sets. The XGBoost (eXtreme Gradient Boosting) classifier algorithm which is a supervised learning technique was used to build the model on the training set for the binary and tertiary classification models. The grid search method was used where multiple model parameters were tuned with a five-fold cross validation to predict the severity pf LBP. A random forest regression model was also built to see which of the models performed the best. Model performance between a linear regression model and the random forest model was compared and the random forest was found to be better with an RMSE of 4.7 vs 11.8.

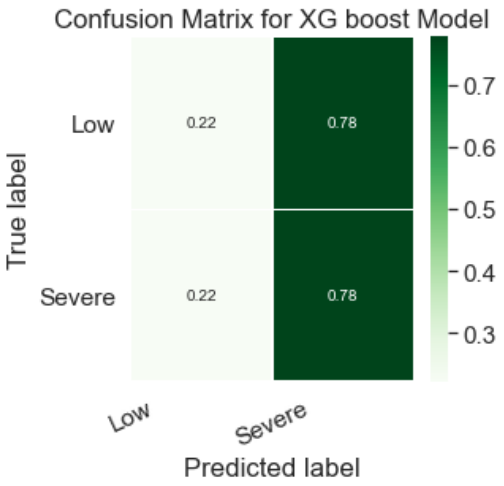
Results

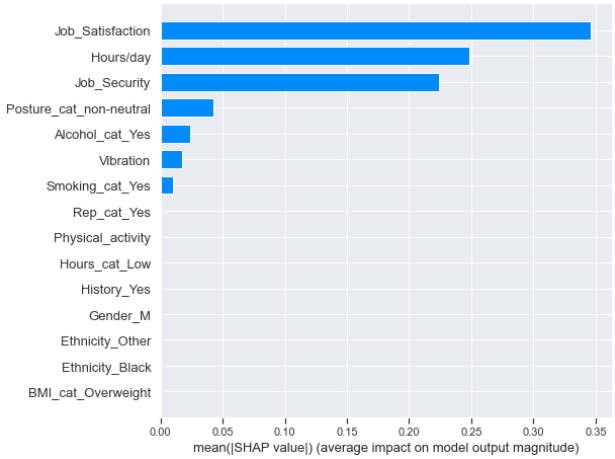
Tertiary Classification Model – An accuracy score of 0.61 was obtained. The shap plot below show the features and their contribution to the model's output. Job Satisfaction seems to be the most important feature affecting the classification decision with a positive correlation, followed by no of years.

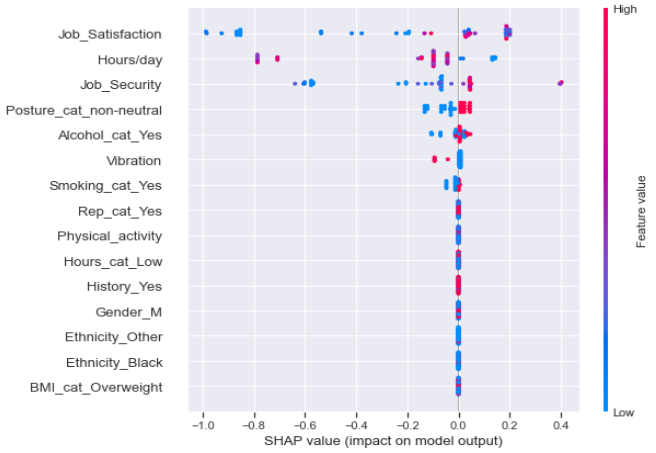




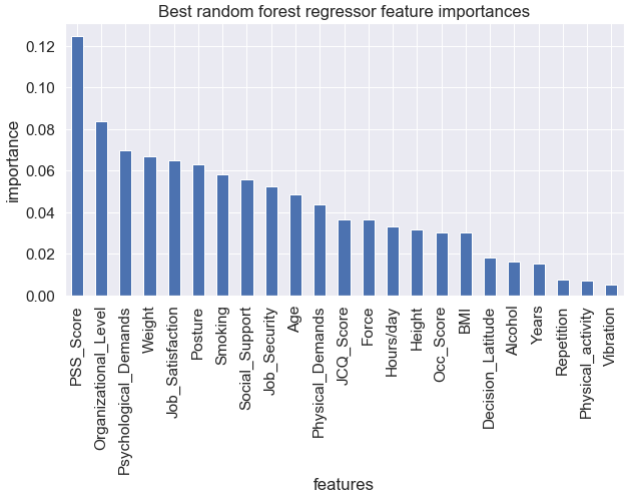
Binary Classification Model – An accuracy score of 0.50 was obtained. The summary and shap plots below show the features and their contribution to the model's output. Job Satisfaction seems to be the most important feature affecting the classification decision with a negative correlation, followed by Hours/day.







Regression Model – An RMSE value of 5.6 was obtained.



Discussions & Conclusions

The tertiary classification model seems to be the best predictive model. Job Satisfaction with a negative correlation seems to be an important indicator present in both the classification models whereas PSS was seen to be most important in the regression model. This shows that psychosocial factors related to job satisfaction and perceived stress would need to be considered when designing preventive measures for high LBP severity. One limitation of this project the limited sample size. A larger sample should probably be considered in model building to obtain a more efficient model.