Capstone Project Documentation

# Problem Statement

## What is the problem of the opportunity that the project is investigating?

* Provide a reliable tool to gauge when to refer patients to the appropriate health professional (chiropractor/physio) to ensure optimal patient outcome
* Provide a screening tool for private health insurance companies to see which customers are more at risk of back pain and future injuries and complications

## Why is this problem valuable to address?

* Low back pain is the leading cause of disability globally, and expected to continue to increase due to ageing and obesity
* It is growing rapidly while being poorly understood, and mismanaged medically
* Australia spends $4.8 billion per year on management of LBP
* It reduces Australia’s GDP by $3.2 billion per annum 🡪 it is the number one condition keeping older Australians out of the workforce

## What is the current state?

* Due to ageing and sedentary lifestyles, low back pain is rapidly increasing
* At the same time, the education is inadequate and therefore there is inefficient medical mismanagement causing a waste of financial and medical resources

## What is the desired state?

* Correctly identify the right patients and refer to the right treatments
* This helps to greatly improve patient outcomes while reducing/spreading the burden of cost in different healthcare areas

## Has this problem been addressed by other research projects?

* A lot of studies focus on measuring the efficacy of specific treatment modalities to see which ones are effective and which ones are not
* Other research focuses on diagnostics using expensive imaging technology such as MRI, which may not be readily available of financially justified for non-urgent back pain diagnosing

# Industry / Domain

## What is the industry / domain?

* Healthcare

## What is the current state of this industry?

* Traditionally, low back pain was predominately treated with analgesics and prescribed bed rest – which is now considered detrimental to patient recovery
* Australian government introduced a health initiative called Chronic Disease Management (CDC), or formerly known as Enhanced Primary Care (EPC) for GP’s to coordinate health care with other health professionals

# Stakeholders

## Who are the stakeholders?

* Medical doctors, health insurance companies, health policymakers, allied health professionals

## Why do they care about this problem?

* Medical doctors: ensure they are treated by more specialised health professionals, and freeing their resources
* Health insurance companies: identify patients at risk of injury for financial and occupational purposes
* Health policymakers: assist in allocating resources in identifying low back pain to reduce GDP loss due to disability
* Allied health care professionals: increased utilisation of their services and improve patient outcomes

## What are the stakeholders’ expectations?

* Able to use this modelling to improve decision makings in terms of policy, and increase confidence in appropriate patient referral and management

# Business Question

## What is the main business question that needs to be answered?

* How can we use low-cost imaging to identify patients at risk of disc herniation or spondylolisthesis?

## What is the business value of answering this question?

* Australia loses $3.2 billion dollars from LBP
* Australia spends over $4.8 on LBP treatment
* Physiotherapist treatment costs $6,063 per episode of care and gain $32,155 in Quality-Adjusted Life-Year compared to the Australian government guidance of $215,875

## What is the required accuracy? What are the implications of false positives or false negatives?

* The accuracy does not need to be extremely high for usefulness, nor would it cause adverse outcomes
* The implications of false positives would not create adverse affects

# Data Question

## What is the question that needs to be answered?

* Which spinal structural parameters can we use to identify patients at risk of disc herniation or spondylolisthesis?

## What is the data required to answer the question?

* Dataset has 310 observations with 13 columns containing specific spinal biomechanical attributes, derived from the shape and orientation of the lumbar spine and pelvis

# Data

## Where was the data sourced?

* https://archive.ics.uci.edu/ml/datasets/Vertebral+Column

## What is the volume and attributes of the data?

* 310 observations with 12 columns of spinal shape and orientation, and each observation is attributed to the target variable of: normal/abnormal.
* Abnormal refers to patients at risk of developing spondylolisthesis and disc herniation

## How reliable is the data?

* Could be improved with explanation of some parameters

## What is the quality of the raw data?

* Very good, no missing null data

## How was this data generated?

* Collected in a hospital setting

## Is this data available on an ongoing basis?

* This data has not been updated since

# Data Science Process

# Data Analysis

## What was the highlights of the EDA?

* There is significant correlation between pelvic incidence and lumbar lordosis angle, pelvic tilt, sacral slope
* Those at risk of disc herniation and spondylolisthesis generally tend to have higher: pelvic incidence, pelvic tilt, lumbar lordosis angle, pelvic radius

## Is the pipeline reusable?

* Yes, the same pipeline can be used to generate future classifications

# Modelling

## What are the main features used?

* All features were used

## Did you find any interesting interactions between features?

* Those related together by region anatomically had a strong correlation with each other

## Is there a subset of features that would get a significant portion of your final performance? Which features?

* Ideally the main features would be pelvic incidence, sacral slope, pelvis radius, pelvic tilt, lumbar lordosis angle, and degree spondylolisthesis

## How did you select features?

* Decision Tree Classifier

## Which feature engineering techniques are used?

* Decision Tree Classifier

## What are the models used?

## Support vector machine, gradient boosing, random forest, logisitic regression

## What are the model performance metrics?

* Precision, recall, f1-score

## Which model was selected?

## Logistic regression

# Outcomes

## What are the main findings and conclusions of the data science process?

* The user can used this model to correctly classify a patient at risk of spondylolisthesis or disc hernation to a fairly good degree of confidence

# Implementation

## What are the considerations for implementing the model in production?

## The model cannot use a radiograph taken directly, it must be measured on which the features are based upon

# Data answer

## Was the data question answered satisfactorily?

* Yes

## What is the confidence level in the data answer?

* Fairly good, but can be improved by removing certain features

# Business answer

## Was the business answered satisfactorily?

* Yes this is model that can be used by stakeholders

## What is the confidence level in the business answer?

* The pipeline is ready to be used, and utilisation and implementation of the model can be increased by strengthening the data acquisition portion

# Response to stakeholders

## What are the overall messages and recommendations to the stakeholders?

* This model will enable timely classification of patients under the right setting, thereby increasing utilisation of the right health services at the right time, offsetting the cost to the public health system

# End-to-end solution

## What is the overall end-to-end solution to use the model developed in the project?

* The ultimate goal is for radiograph of the patient to be taken, and computer vision implementation will automatically and objectively measure the radiographs, and then it will be feed directly into the model for classification

# References

## Where are the data and code used in the project?

https://github.com/Peter-L20/Institute-of-Data