Project 3: Football prediction

Background

The English Premier League is the most-watched sports league in the world, and accurately forecasting the results is of great importance to coaches, fans and bookmakers. The data set for this project comprises over 60 statistics collected on every Premier League game (380 per season).

Your aim

Devise a statistical model which accurately forecasts the result of future football matches. Can your model forecast results in other football leagues?

Your data

The data set can be downloaded from http://www.football-data.co.uk/data.php

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used as well as results indicating how accurate the forecasts are. The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second- year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Colon cancer

Background

The importance of increasing the survival of cancer patients cannot be overstated and thus accurately identifying factors that are associated with survival is a topic of great interest for the statistics community. The data for this project contain survival information on 15564 patients diagnosed with cancer of the colon in a Northern European Country between 1975 and 1994.

Your aim

Devise a statistical model which adequately models the survival time of patients as a function of possible prognostic variables. Can you identify the most important of these variables in terms of prognosis for survival? Can you predict the survival for a given individual when the value of its prognostic variables is known?

Your data

The data set can be downloaded from the module's moodle page.

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used. Possible things to consider are censored and missing data and how they may (or not) affect the analysis.

The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second-year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Ranking tennis players

Background

Sports rankings are designed for various purposes. Professional tennis players in particular aspire to hold the top ranking position; reaching the top 100 can be seen as a significant career milestone. The data set for this project comprises match results of the Association of Tennis Professionals (ATP) tour since 2000.

Your aim

Devise a statistical model which ranks male tennis players. How does the ranking of players change for different types of tournaments? Can your model take into account match characteristics? How do your model's rankings compare to published rankings of tennis players?

Your data

The data set can be downloaded from http://tennis-data.co.uk/

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used. The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second- year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Credit Scoring

Background

Credit scoring is a tool used by lenders (e.g. banks) to help decide whether one qualifies for a particular credit card, loan, mortgage or service. Credit score is a number everyone over the UK over the age of 18 is given to indicate their credit worthiness. Each lender uses a different method to calculate their own credit scores and it is related to the probability of default. The data set for this project provides information on 150,000 borrowers.

Your aim

Develop a statistical model / algorithm which accurately predicts the probability of a default on a loan based on data available to investors. Which attributes are more useful in making such predictions?

Your data

The data set can be downloaded from the module's Moodle page.

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used as well as results indicating how accurate the predictions are. The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second- year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Ebola outbreak

Background

The 2014 Ebola virus disease outbreak in West Africa was the largest outbreak of the genus Ebolavirus in history. A real-time analysis of the numbers of infected individuals and deaths due to Ebola could provide helpful information for public health policy. The data set for this project consists of the number of cases and deaths as reported in the World Health Organization (WHO) situation reports for Guinea, Liberia, and Sierra Leone during the 2014 outbreak.

Your aim

Develop a statistical model that accurately describes the course of the outbreak over its full duration. Does your model provide estimates of transmission rates and other related parameters? Can you estimate the average number of secondary infections generated by an infectious individual? Does your model predict the course of the outbreak during 2015 by utilising the data available up to the end of 2014 only? Are the dynamics of the disease common to the three different regions? **Your data**

The data set can be downloaded from the WHO situation reports:

https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/case-counts.html

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the models and methods used as well as results indicating how accurate the forecasts are. The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second- year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Peer-to-Peer lending

Background

Lending Club (LC) is the world's largest peer-to-peer lending platform. A borrower can apply for a loan, and if accepted by LC, their loan gets listed in the marketplace. An investor can browse loans in the marketplace, and invest in individual loans at their discretion. The data set for this project comprises information on the loans issued by LC until the end of 2018.

Your aim

Develop a statistical model which adequately predicts the probability of a default on a loan based on data available to investors. Which attributes are more useful in making such predictions? Is there any association between the predicted probability of default and interest rate?

Your data

The data set can be downloaded from `https://tinyurl.com/wtcca5a`

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used as well as results indicating how accurate the predictions are.

The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second-year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: Climate forecast

Background

Long-term weather forecasting is important for planning, to mitigate the effects of climate change and extreme weather events. The data set for this project comprises information on temperature, rainfall and sun light collected daily at over 100 weather stations in Australia.

Your aim

Devise a statistical model which adequately forecasts the weather in the long term. Can your model forecast the weather at other locations in Australia? Can your model forecast how often extreme weather events will occur?

Your data

The data set can be downloaded from http://www.bom.gov.au/climate/data/

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used as well as results indicating how accurate the forecasts are.

The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second-year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.

Project 3: River flows

Background

Accurate forecasting of river flows is important for planning projects such as flood defences. The data set for this project comprises the daily flow of water as measured at around 1500 locations in the UK, together with data on the river catchment.

Your aim

Devise a statistical model which accurately forecasts the flow of rivers. Can your model forecast river flow at other locations in the UK? Can your model forecast how often a river flow of certain strength will be exceeded?

Your data

The data set can be downloaded from http://nrfa.ceh.ac.uk/data/search

Your task

The group should submit a written report addressing the aim outlined above. It should include sufficient theoretical justification for the methods and models used as well as results indicating how accurate the forecasts are. The report should be approximately 25 sides of text (30 sides max) in a 12pt size font, single-spaced, and with margins of at least 2cm on all sides (title page, abstract, references and appendices are not included in the page count). You should assume your audience are third-year mathematics students who have studied first and second- year modules in probability and statistics. The group should also give an in-person 15-minute oral presentation supported by the use of slides during the session on Friday the 15th of March and be prepared to answer questions.

Submission requirements

Each group must submit their report, presentation slides and a single project submission form via Moodle by 16:00 on Thursday 14th of March in **pdf** format – please check the module's page for details.