Sports performance analysis Research
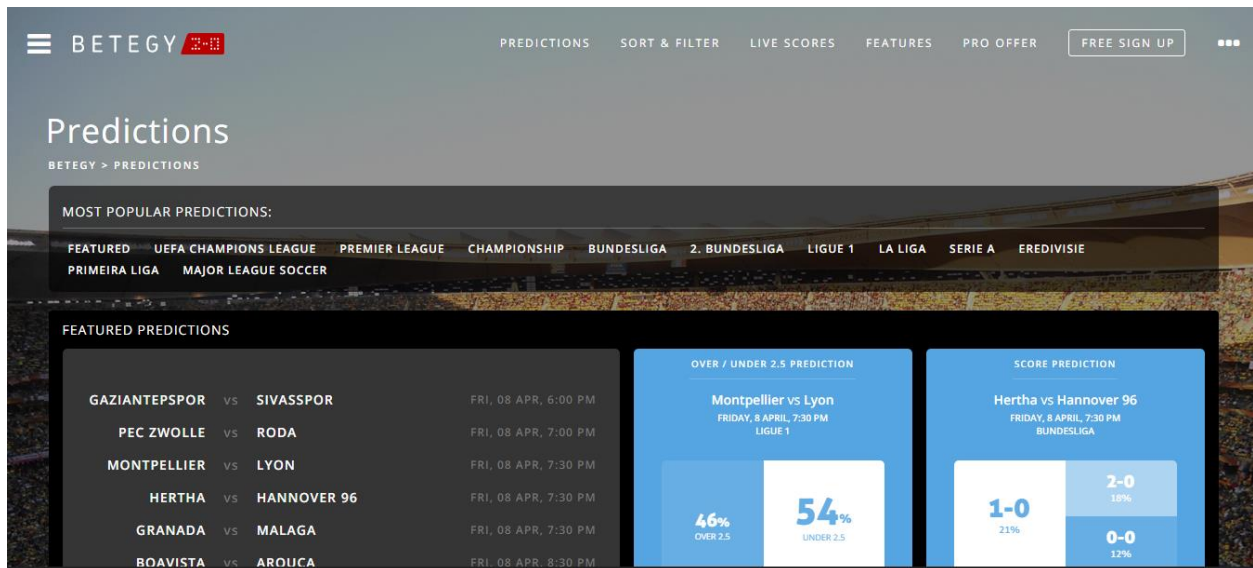
Author: Branko Chomic

Date: 14/04/2016

Table of Contents

Introduction

The purpose of this project is to build a web application which could help people with decision making when placing bets in most popular sports such as football, basketball, American football and hockey and it would predict outcomes by analyzing statistics and creating training and testing data sets from them to be able to predict outcomes using machine learning algorithms. The content of the website should be updated automatically and it should run on server/cloud . The web application should also contain content such as latest scores, tables , upcoming fixtures, statistics for each team and live scores.

Similar applications:

1. Betegy

[1]Betegy is web application which is available for phones and desktop users providing users with predictions for more than 22 leagues in football.

Features:

- 1x2 prediction card
- Score prediction card
- Over 1.5 / Under 1.5 prediction card
- Over 2.5 / Under 2.5 prediction card
- Over 3.5 / Under 3.5 prediction card
- Both Team to Score prediction card
- Handicap prediction card
- Algorithm Recommendation card

Predictions:

Algorithm powered by Modified Dixon & Coles model Uses Neural Network functionality to predict results

Analyses Stats (points, goals, game scores, league position, struggles, players' characteristics)

Examines team dynamics (current form, progression, regression, motivation)

Accounts for the latest football news & data (injuries, weather, transfers, manager & coach impact)

My opinion:

I was using this website for few weeks and I think it is the best one I came across so far I really like the user friendly design of the website as well as that they provide many different predictions like (both team to score, score predictions, handicap predictions) which gives users a lot of options. The only thing I didn't like in this web application is the fact that it only supports predictions for one sport.

2. Pick for win
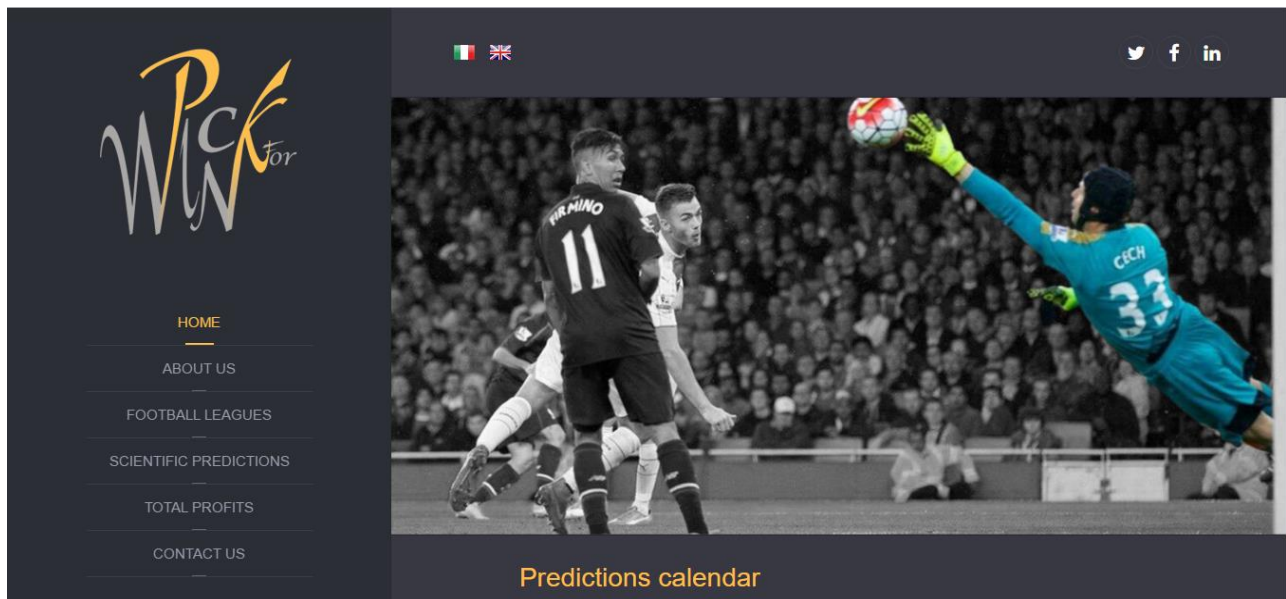
Features:

[2]Predictions for 2 sports including basketball and football with variety of leagues to choose from.

It predicts data also by analyzing statistics of each team and based on that they give probability percentage of every possible outcome(home win, draw or away win). They also show odds for each game and what outcome should user pick when placing a bet.

My opinion:

I think this web app has very interesting and simple design which is very easy to use it supports predictions for top leagues in football. Good thing about this web app is that its is completely free for all users the thing I didn't like is that it doesn't display statistics of each team which the prediction are based on.

3. 7M Sports

[3]7M sports is great web application for users who likes sports. This website includes predictions of 4 different sports(football, basketball, tennis and baseball) it is also available as an application for android and ios users.

Features:

- Prediction for four sports(football, basketball, tennis and baseball)
- Displaying latest news
- Videos from press conferences, highlights of the games
- Live scores
- IOS and android application available

My opinion:

The things that I like about this web application is that it provides more than one sport to choose from also I like the fact that it show latest soccer news , live scores and betting odds which are very useful features to have in a sport prediction web application.

Machine learning

Introduction

[4]Machine learning is a method of data analysis that is analyzing model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

Types of machine learning:

1.Supervised learning

Is a machine task of inferring a function from supervised training data set. Training data set always contains set of training examples, each example contains a input object and a desired output value. The goal of supervised learning is to build an artificial system that can learn the mapping between the input and the output, and can predict the output of the system given new inputs. If the output takes a finite set of discrete values that indicate the class labels of the input, the learned mapping leads to the classification of the input data. If the output takes continuous values, it leads to a regression of the input. The input-output relationship information is frequently represented with learning-model parameters. When these parameters are not directly available from training samples, a learning system needs to go through an estimation process to obtain these parameters. Supervised learning has a lot of usefully algorithms which includes the likes of (Neural networks, Nearest neighbor, Random Forests, Bayesian) and many more.

2. Unsupervised learning

Is a type of machine learning algorithm which is used for drawing of the interfaces and datasets consisting of input data without labeled responses. The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.

Conclusion

For my project supervised learning is more suitable because I will need to create training data set with desired output and i will need supervised learning algorithms to be able to learn and predict the correct output.

Machine learning algorithms(Supervised learning)

Artificial neural networks

In machine learning neural networks are models which are inspired by biological neural networks and are used to estimate functions that are depending on large amount of inputs and are unknown. Neural networks consists of neurons and each neuron have a connection to next neuron. Each connection has a numeric weight which can be tuned making neural nets more adaptive to different inputs.

Multi-layer perceptron

Is a neural network model that maps sets of appropriate inputs. Multilayer perceptron consists of input nodes which are connected to each other with layers. It uses back propagation supervised learning technique. Multi layer perceptron is widely used for predicting sport matches outcomes as well as face recognition and many other different problems.

Linear regression

Linear regression is an algorithm for modeling a relationship between scalar variable y and variables x. The relationships are modeled using linear predictor function whose unknown outputs are predicted based on data these models are called linear models.

Linear regression has many uses:

If the goal is prognostication, or forecasting, or error reduction, linear regression can be acclimated to fit a predictive model to an observed data set of y and X values. After developing such a model, if an adscititious value of X is then given without its accompanying value of y, the fitted model can be acclimated to make a prognostication of the value of y.

Given a variable y and a number of variables X1, ..., Xp that may be cognate to y, linear regression analysis can be applied to quantify the vigor of the relationship between y and the Xj, to assess which Xj may have no relationship with y at all, and to identify which subsets of the Xj contain redundant information about y.

Conclusion

For this project usage of more supervised learning algorithms will be used and tested to see which algorithm is most suitable for this kind of problem. The algorithms that will be tested are linear regression, Bayesian network, multi-layer perceptron.

Target language?

Python

[5]Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C. The language provides constructs intended to enable clear programs on both a small and large scale. Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library.

Advantages

Python code has to be strictly indented but the indentation helps in much cleaner code(readable)

2. High level data structures are list, directory are well suited in python.

3. Easy to write, easy to read and easy to understand.

Disadvantages

The documentation is not good as some of the other languages like php or java that have strong corporate backing.

Python is interpreted language and is very slow when compared to other languages like java or c++.

Python is not very good for memory intensive tasks

Java

[6]Java programming language was pristinely developed by Sun Microsystems which was initiated by James Gosling and relinquished in 1995 as core component of Sun Microsystems' Java platform (Java 1.0 [J2SE]).

The latest relinquishment of the Java Standard Edition is Java SE 8. With the advancement of Java and its widespread popularity, multiple configurations were built to suite sundry types of platforms. Ex: J2EE for Enterprise Applications, J2ME for Mobile Applications.

The incipient J2 versions were renamed as Java SE, Java EE and Java ME respectively. Java is ensured to be indicted once, it can run anywhere.

Advantages

Object Oriented: In Java, everything is an Object. Java can be facilely elongated since it is predicated on the Object model.

Platform independent: Unlike many other programming languages including C and C++, when Java is compiled, it is not compiled into platform categorical machine, rather into platform independent byte code. This byte code is distributed over the web and interpreted by virtual Machine (JVM) on whichever platform it is being run.

Simple: Java is designed to be facile to learn. If you understand the fundamental concept of OOP Java would be facile to master.

Secure

With Java's secure feature it enables to develop virus-free, tamper-free systems. Authentication techniques are predicated on public-key encryption.

Architectural-neutral:

Java compiler engenders an architecture-neutral object file format which makes the compiled code to be executable on many processors, with the presence of Java runtime system.

Portable

 Being architectural-neutral and having no implementation dependent aspects of the designation makes Java portable. Compiler in Java is indited in ANSI C with an unsullied portability boundary which is a POSIX subset.

Robust:

Java makes an effort to eliminate error prone situations by accentuating mainly on compile time error checking and runtime checking.

Multithreaded

 With Java's multithreaded feature it is possible to indite programs that can do many tasks simultaneously. This design feature sanctions developers to construct smoothly running interactive applications.

Interpreted

Java byte code is translated on the fly to native machine ordinant dictations and is not stored anywhere. The development process is more rapid and analytical since the linking is an incremental and light weight process.

High Performance

With the utilization of Just-In-Time compilers, Java enables high performance.

Distributed

Java is designed for the distributed environment of the cyber world.

Dynamic

Java is considered to be more dynamic than C or C++ since it is designed to acclimate to an evolving environment. Java programs can carry extensive amount of run-time information that can be acclimated to verify and resolve accesses to objects on run-time.

Disadvantages

Performance

Java can be perceived as significantly more gradual and more recollection-consuming than natively compiled languages such as C or C++.

Look and feel

The default look and feel of GUI applications indicted in Java utilizing the Swing toolkit is very different from native applications. It is possible to designate a different look and feel through the pluggable look and feel system of Swing. Sun is working in this area and we can visually perceive some great technologies from Sun in near future.

Single-paradigm language

Java is predominantly a single-paradigm language. However, with the advisement of static imports in Java 5.0 the procedural paradigm is better accommodated than in earlier versions of Java.

My Decision:

For my project I decided to choose java language because it is faster than python and it have a lot of very useful libraries for machine learning and parsing web page content.

How will I get content for my website?

Getting the content for my website is the most important part of this project. The content should be updated regularly to provide users with fresh content. There are few options that are available for me to get the data I need.

Sports API

Sports API are one of the option to obtain content that will be needed for this project. The disadvantage with this approach is that it is mostly not open source and the once that are doesn't have enough of content or the content is out of date.

Web scrapping

Web scrapping is a technique of extracting large amount of data from websites where data is extracted and saved on a computer or to a database.

The data that websites provides can only be displayed by web clients. Most of the websites don't offer the functionality to save a data from their website to your computer. If that's the case than the only option is to manually copy and paste data to your local computer it's a job that can take many hours sometimes even days before it is complete. Web scrapping is a technique of making this process automatically so instead of manually copying and pasting data web scrapping software will perform the same tasks very fast.

There is a lot of open source software and java libraries available for web scrapping.

Advantages of using web scrapping instead of sport API

Availability

[7]Most of the websites provides API today but most of the time there are many limitations on the data that are available through the API. Also while websites are concern about the uptime and availability of the websites because it caters to their most important customers, unavailability and downtime for the API can sometimes go unnoticed for days.

Up-to-date

APIs are used to get updated very slowly because they are normally at the bottom of the priority list. The data served by the API may sometimes be old too. Instead, when you scrape the content off the website, you get what you see. You can easily verify this data.

Better Structure

Navigating through a badly structured API can be very tedious and time consuming. You might have to make dozens of queries before getting to the actual data that you need.

Websites nowadays have a better structure than they have ever had. With every site wanting to be XHTML validated in order to fare better rankings on search engines, the structure of the websites today is clean and easy to scrape. Increasing use of JSON, JSONP, XML and Micro data in the systems have further structured the data used on websites. In the past it was very hard to scrap data from websites because of using regex although the regex is not completely removed there are libraries and tools that make web scrapping easier and faster.

Legal Issues

[11]Many websites this days are not available for scraping content they have mention it in their terms of services that scraping is not allowed. In such a case the best thing to do is to stay way from those websites. The scraping public data might not be an criminal offense, if the source wants you to stop, they can get

injunction against you. With advancement in technology, it has become a war between people who tries to scrap the content and those who tries to block such requests. There are no proper laws which defines the scope and legality pertaining to web scraping.

Conclusion

The advantages stated above are the reason for me to choose web scrapping rather than API as my content on the web application should be automatically updated every day to keep the content and the statistics for prediction up to date.

Web services

1.Apache tomcat

[8]Apache Tomcat™ is an exhibit source software implementation of the Java Servlet, JavaServer Pages, Java Expression Language and Java WebSocket technologies. The Java Servlet, JavaServer Pages, Java Expression Language and Java WebSocket specifications are spread under the Java Community Process.

Apache Tomcat is extended in an bring to light and participatory environment and declared publicly under the Apache License explanation 2. Apache Tomcat is sealed to be a synergy of the best-of-breed developers from far and wide the world. We recall you to receive in this bring to light development project. To commemorate more roughly getting preoccupied, gat along well here.

Apache Tomcat powers beyond number large-scale, mission-critical internet applications adjacent a antithetical range of industries and organizations.

Advantages

Lightweight:

Tomcat is incredibly lightweight application. Tomcat offers the most basic functionalities that are needed to run a server and that means that it loads very quickly and redeploy times are fast as well when compared to other servers. The lightweight nature of apache tomcat means that the development cycle is faster as well.

Open source

For me, open-source always counts as a acquire victory. Tomcat's free, and the source code for the server is yarely available to anyone who'd care to download it. What this denotes is that – postulating you're inclined to tinker with the moving components of your server – you've got an incredible degree of liberation insofar as what you optate to do with a Tomcat installation.

Flexible

Thanks to its lightweight nature and a suite of extensive, built-in customization options, Tomcat is quite flexible. You can run it in virtually any fashion you optate, and it'll still work as intended. The fact that it's open-source avails as well, since you can tweak it to fit your needs.

Security

Many organizations decide to position their Tomcat installation abaft an extra firewall, accessible only from the Apache installation. Depending on how you install apache tomcat it can provide extra security layer which is always a good thing.

2.Amazon services

[9]Amazon Web Accommodations provides a highly reliable, scalable, low-cost infrastructure platform in the cloud that powers hundreds of thousands of businesses in 190 countries around the world. With data center locations in the U.S., Europe, Brazil, Singapore, Japan, and Australia, customers across all industries are capitalizing on the following benefits:

Advantages:

Low Cost

AWS offers low, pay-as-you-go pricing with no up-front expenses or long-term commitments.

Limberness and Instant Elasticity

AWS provides a massive ecumenical cloud infrastructure that sanctions you to expeditiously innovate, experiment and iterate. In lieu of waiting weeks or months for hardware, you can instantly deploy incipient applications, instantly scale up as your workload grows, and instantly scale down predicated on demand.

Open and Flexible

AWS is a language and operating system agnostic platform. You optate the development platform or programming model that makes the most sense for your business.

Secure

AWS is a secure, durable technology platform with industry-apperceived certifications and audits: PCI DSS Level 1, ISO 27001, FISMA Moderate, Fed RAMP, HIPAA, and SOC 1 (formerly referred to as SAS 70 and/or SSAE 16) and SOC 2 audit reports.

Google App Engine [10]

● Leading Java and Python PaaS public cloud

● Infrastructure similar to the one driving GMail and Google Docs

● App engine has a completely free version

● Provides CPU / bandwidth / storage capable to serve 5 000 000 page views per month

● Instant registration

"Google App Engine launched in the spring of 2008. It was different from most other cloud systems back in the day because it was neither IaaS (Infrastructure-as-a-Service, e.g., Amazon EC2) nor SaaS (Software-as-a-Service, e.g., Salesforce). It was something in-between and ushered in the era of PaaS (Platform-as-a-Service). Instead of a fixed application (SaaS) or raw hardware (IaaS).App Engine manages your infrastructure for users. Furthermore, it provides

a development platform... users get to create apps, not used the one provided by the cloud vendor, and it leverages the infrastructure as a hosting platform."

Advantages

Will automatically handle and balance all instances and data centers for web applications

Provides support for 4 popular programming languages: PHP, Python, Java, and Go

Will automatically scale data storage or instances to meet the needs of the programmer

Provides App Engine SDK to produce applications locally

App Engine can be integrated into multiple IDEs

Has a secured Internet Infrastructure to store code and application data safely

Disadvantages

If Google servers go down, applications will not run properly

Must be careful to read the long list of policies. Violation of Google AdSense policies are common and have resulted in bans

Everything is handled by Google: company data information and pricing changes are all in Google's hands.

Not flexible for future opt-out. If you leave GAE, it is not possible to port your app and you will need to develop your application from scratch again

Conclusion

Amazon services is flexible and allows users to create virtual machines for operating systems like Linux and windows which is something I will need in development of my web application as I intend to store and run my code online. Amazon services are also easy to use as they provide a lot of tutorials for beginners .These reasons made me decide to go with amazon services over other web services I researched.

References

[1] Name: Betegy, Author: Betegy, accessed 18 October 2015

https://betegy.com/features

[2] Name: Pickforwin, Author: Pickforwin, accessed 19 October 2015

http://www.pickforwin.com/en/about.html

[3] Name: 7M sports, Author: 7m sport.com, accessed 18 October 2015

http://www.7msport.com/

[4] Name: machine learning, Author: sas, accessed 18 October 2015

http://www.sas.com/en_us/insights/analytics/machine-learning.html

 [5] Name: Python, Author: Wikipedia, accessed 15 October 2015

https://en.wikipedia.org/wiki/Python_(programming_language)

[6] Name: Java, Author: Wikipedia, accessed 2 October 2015

https://en.wikipedia.org/wiki/Java_(programming_language)

[7] Name: Web scrapper vs API Author: Grepsr, accessed 21 October 2015

https://www.grepsr.com/web-scraping-vs-api/

[8] Name: Apache Tomcat Author: Apache , accessed 8 October 2015

http://tomcat.apache.org/

[9] Name: Amazon services : Amazon, accessed 18 November 2015

https://aws.amazon.com/about-aws/

[10] Name: Google App engine Author: Google, accessed 5 October 2015

https://cloud.google.com/appengine/docs

 [11] Name: Web scrapping legal issues Author: tonido, accessed 25 December 2015

http://www.tonido.com/blog/index.php/2013/12/28/web-scraping-and-legal-issues/#.Vw5UGTArLIU