Project Config Manual

Configuration manual for the research project submitted as part of the MSC in Data Analytics

Oisin Wiseman (14108399) 8/22/16 MSC in Data Analytics

Table of Contents

Introduction:

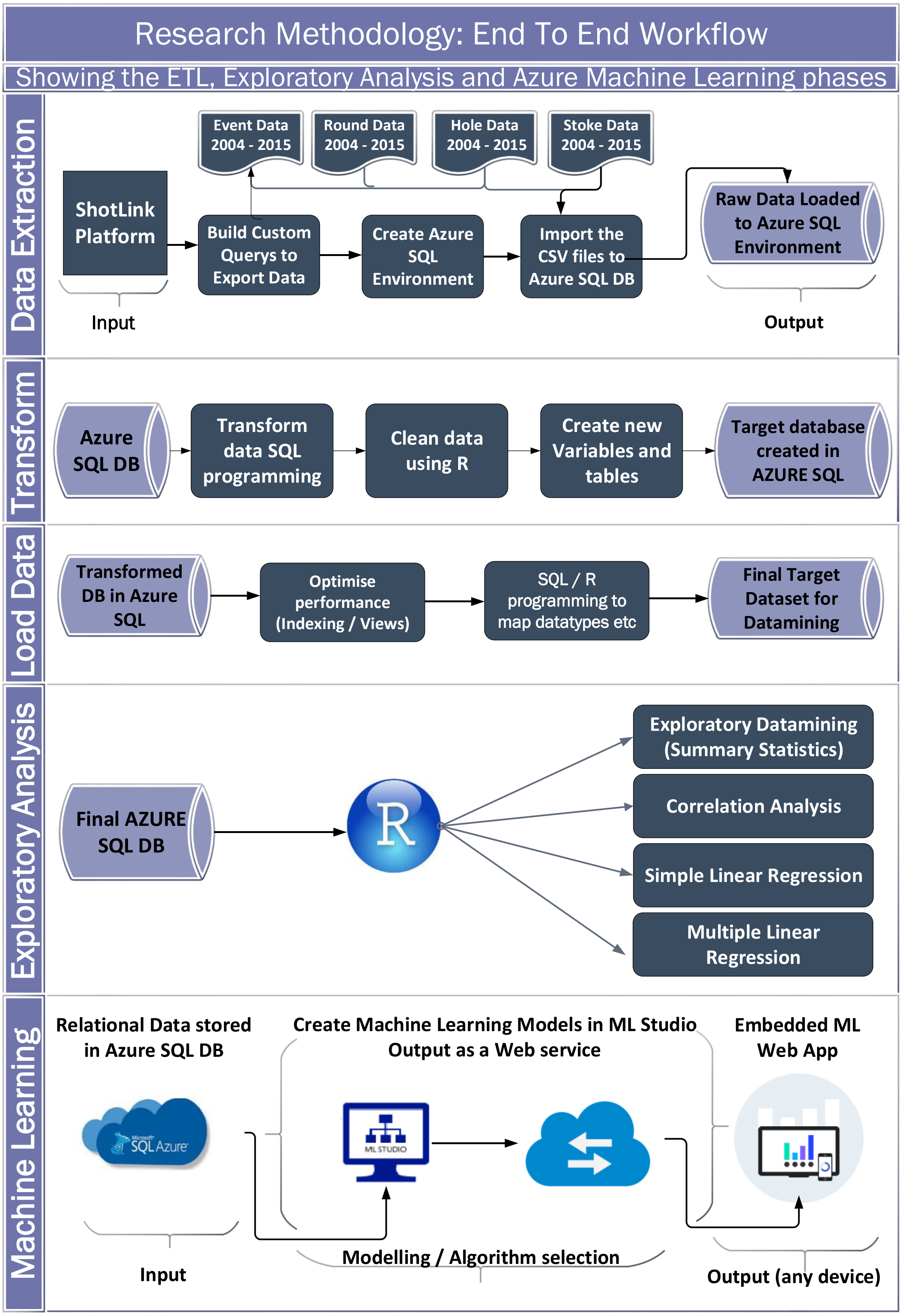
This manual is to accompany the research paper submitted as part of the MSc. In Data Analytics entitled 'Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour'. The data in this manual discusses the technology used, how it was applied and talks through in detail the key pieces related to the paper so they can be reproduced within the college.

Technology

The following software and technical resources were used through this research project and more detail will be provided throughout this document.

- The ShotLink Platform: The ShotLink system is accessible through a secure website https://stats.pgatourhq.com/ for members of the PGA TOUR. Access for this research was gained by applying for permission in writing to the PGA TOUR and signing a Non-Disclosure Agreement(NDA). A copy of this NDA is included with the artefact.
- Microsoft Azure SQL: This is a database as a service a cloud offering from Microsoft. It was envisioned that all work for this project would be carried out in the Azure cloud platform. All the target data was stored in Azure SQL for this research. This was used in conjunction with Microsoft SQL Server 2016.
- Microsoft Azure Machine Learning: A cloud-based predictive analytics service from Microsoft. It provides the platform and tools to easily to model predictive analytics. Learning Studio is the developer environment used to create and publish models as web services. It has built in support for Azure SQL where the target data is stored and also supports custom R code for data transformation. This was used to evaluate and compare the machine learning algorithms and deploy the web services that were used by the final applications.
- Microsoft Power BI Desktop: The majority of the Exploratory Data Analysis was carried out through visualisation using Power BI desktop from Microsoft. Seamless integration with the target data on Azure SQL and enabled powerful yet quick visualisations for help in slicing and dicing the data.
- R – All the statistical analysis described in the final paper was carried out using R and R studio. This includes the exploratory Data Analysis, Correlation analysis, Linear Regression, step regression and relative importance.
- Microsoft SQL Server Management Studio 2016: This was to import the pre-processed data into Azure SQL DB. Transformed the data by setting data-types and memory allocation etc.
- Microsoft Excel 2016 (64 bit): Excel was used to load the exported CSV's from ShotLink and carry out some initial pre-processing. It was also used for the graph in the results section of the paper. Note it was necessary to use the 64-bit version given the size of the CSV's coming from ShotLink.

The workflow on the next page shows where each of these technologies were applied in this research project. This config manual go into more detail on how these were used and what is required to set this project up to reproduce the results.


Research Methodology: End To End Workflow
Showing the ETL, Exploratory Analysis and Azure Machine Learning phases
Data Extraction
Event Data 2004 - 2015
Round Data 2004 - 2015
Hole Data 2004 - 2015
Stoke Data 2004 - 2015
ShotLink Platform
Build Custom Querys to Export Data
Create Azure SQL Environment
Import the CSV files to Azure SQL DB
Raw Data Loaded to Azure SQL Environment
Input
Output
Transform
Azure SQL DB
Transform data SQL programming
Clean data using R
Create new Variables and tables
Target database created in AZURE SQL
Load Data
Transformed DB in Azure SQL
Optimise performance (Indexing / Views)
SQL / R programming to map datatypes etc
Final Target Dataset for Datamining
Exploratory Analysis
Final AZURE SQL DB
R
Exploratory Datamining (Summary Statistics)
Correlation Analysis
Simple Linear Regression
Multiple Linear Regression
Machine Learning
Relational Data stored in Azure SQL DB
Create Machine Learning Models in ML Studio Output as a Web service
Embedded ML Web App
SQL Azure
ML STUDIO
Input
Modelling / Algorithm selection
Output (any device)

The ShotLink Platform:

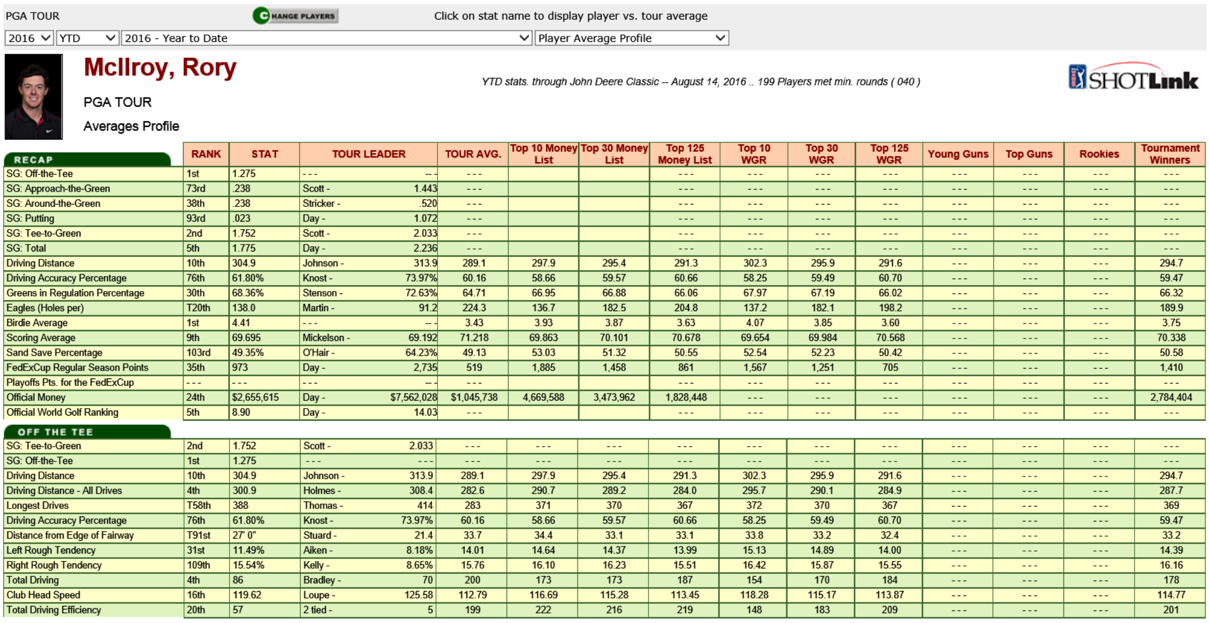
The ShotLink System is the platform used by the PGA TOUR for collecting and disseminating scoring and statistical data on every shot by every player in real-time.

Accessing ShotLink:

Access to ShotLink data is restricted to professional golfers who are members of the PGA tour and others who are granted access as needed. The PGA TOUR provides structured ShotLink datasets to accredited higher educational institutions for research and approved academic purposes. The NCI was not one of their accredited institutions so I applied for access through the email address shotlinkintelligence@pgatourhq.com stating it was for a research project in the NCI. Once I completed an NDA full access to the system was granted. Note this NDA is personal to me and does not cover anyone else in the college. To work with the data access will need to be requested through. shotlinkintelligence@pgatourhq.com

Exporting Data:

The ShotLink system is primarily used by players and coaches to review their progress and statistics throughout the season. Fig 1. is an example of the type of stats that is accessible by members of the tour.



RECAP	RANK	STAT	TOUR LEADER	TOUR AVG.	Top 10 Money List	Top 30 Money List	Top 125 Money List	Top 10 WGR	Top 30 WGR	Top 125 WGR	Young Guns	Top Guns	Rookies	Tournament Winners
SG: Off-the-Tee	1st	1.275	--- --	---			---	---	---	---	---	---	---	---
SG: Approach-the-Green	73rd	.238	Scott - 1.443	---			---	---	---	---	---	---	---	---
SG: Around-the-Green	38th	.238	Stricker - .520	---			---	---	---	---	---	---	---	---
SG: Putting	93rd	.023	Day - 1.072	---			---	---	---	---	---	---	---	---
SG: Tee-to-Green	2nd	1.752	Scott - 2.033	---			---	---	---	---	---	---	---	---
SG: Total	5th	1.775	Day - 2.236	---			---	---	---	---	---	---	---	---
Driving Distance	10th	304.9	Johnson - 313.9	289.1	297.9	295.4	291.3	302.3	295.9	291.6	---	---	---	294.7
Driving Accuracy Percentage	76th	61.80%	Knost - 73.97%	60.16	58.66	59.57	60.66	58.25	59.49	60.70	---	---	---	59.47
Greens in Regulation Percentage	30th	68.36%	Stenson - 72.63%	64.71	66.95	66.88	66.06	67.97	67.19	66.02	---	---	---	66.32
Eagles (Holes per)	T20th	138.0	Martin - 91.2	224.3	136.7	182.5	204.8	137.2	182.1	198.2	---	---	---	189.9
Birdie Average	1st	4.41	--- --	3.43	3.93	3.87	3.63	4.07	3.85	3.60	---	---	---	3.75
Scoring Average	9th	69.695	Mickelson - 69.192	71.218	69.863	70.101	70.678	69.654	69.984	70.568	---	---	---	70.338
Sand Save Percentage	103rd	49.35%	O'Hair - 64.23%	49.13	53.03	51.32	50.55	52.54	52.23	50.42	---	---	---	50.58
FedExCup Regular Season Points	35th	973	Day - 2,735	519	1,885	1,458	861	1,567	1,251	705	---	---	---	1,410
Playoffs Pts. for the FedExCup	---	---	--- --	---			---	---	---	---	---	---	---	---
Official Money	24th	$2,655,615	Day - $7,562,028	$1,045,738	4,669,588	3,473,962	1,828,448	---	---	---	---	---	---	2,784,404
Official World Golf Ranking	5th	8.90	Day - 14.03	---			---	---	---	---	---	---	---	---
OFF THE TEE														
SG: Tee-to-Green	2nd	1.752	Scott - 2.033	---	---	---	---	---	---	---	---	---	---	---
SG: Off-the-Tee	1st	1.275	--- ---	---	---	---	---	---	---	---	---	---	---	---
Driving Distance	10th	304.9	Johnson - 313.9	289.1	297.9	295.4	291.3	302.3	295.9	291.6	---	---	---	294.7
Driving Distance - All Drives	4th	300.9	Holmes - 308.4	282.6	290.7	289.2	284.0	295.7	290.1	284.9	---	---	---	287.7
Longest Drives	T58th	388	Thomas - 414	283	371	370	367	372	370	367	---	---	---	369
Driving Accuracy Percentage	76th	61.80%	Knost - 73.97%	60.16	58.66	59.57	60.66	58.25	59.49	60.70	---	---	---	59.47
Distance from Edge of Fairway	T91st	27' 0"	Stuard - 21.4	33.7	34.4	33.1	33.1	33.8	33.2	32.4	---	---	---	33.2
Left Rough Tendency	31st	11.49%	Aiken - 8.18%	14.01	14.64	14.37	13.99	15.13	14.89	14.00	---	---	---	14.39
Right Rough Tendency	109th	15.54%	Kelly - 8.65%	15.76	16.10	16.23	15.51	16.42	15.87	15.55	---	---	---	16.16
Total Driving	4th	86	Bradley - 70	200	173	173	187	154	170	184	---	---	---	178
Club Head Speed	16th	119.62	Loupe - 125.58	112.79	116.69	115.28	113.45	118.28	115.17	113.87	---	---	---	114.77
Total Driving Efficiency	20th	57	2 tied - 5	199	222	216	219	148	183	209	---	---	---	201

Figure 1: Example of the data available to players in ShotLink

However, for this research we were interested in the historical data for all the years 2004 to 2015. ShotLink proves an option to export data in bulk queries. In order to export the data, goto the 'Tools -> Detail Export' and select the dataset you want and fill out the query as per Fig 2. These are exported as raw CSV you get an email with a download to link. Note that these are quite large the stroke level dataset is over 3.5GB. There are 6 datasets that you can export, for this research we exported four of them as outlined in the paper. These were 'Event Level, 'Round Level', 'Hole Level' and 'Shot Level'.

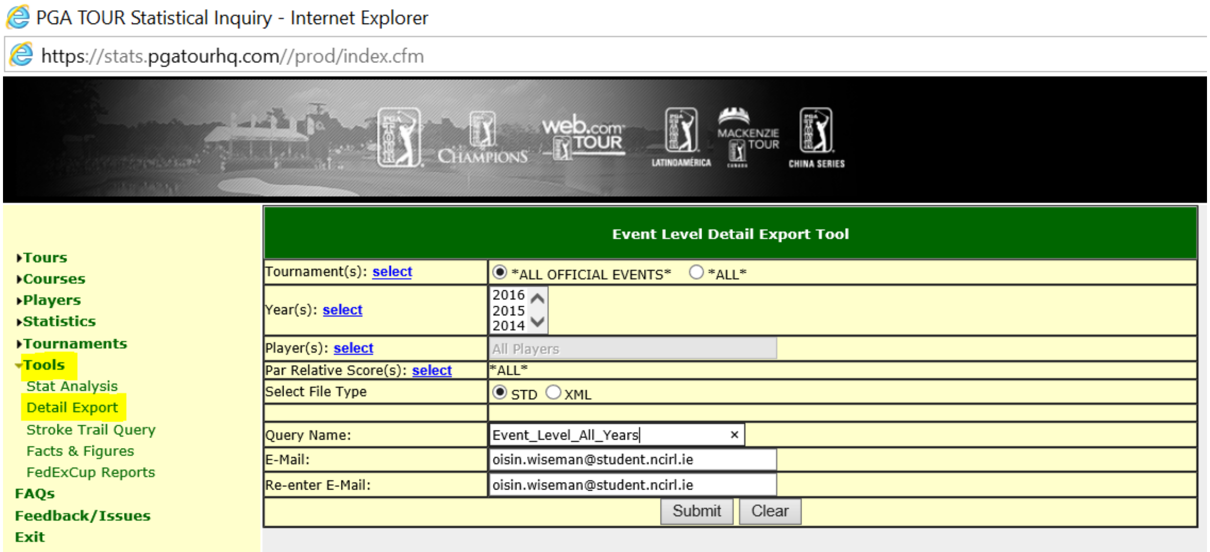


Figure 2: Exporting the Event Data from ShotLink these are exported as CSV

The pre-processed files are included in the electronic submission along with the artefact. Please note though that this data direct from ShotLink and is covered under NDA. In order to work with that data then you need to apply directly to the PGA TOUR.

Once the data is exported from ShotLink as CSV you don't need any further interaction with the ShotLink system. The exported data was then pre-processed in Excel prior to importing into Azure SQL using Microsoft SQL Server Management Studio 2016.

Working the Microsoft Azure Platform:

Azure SQL server was used as the storage platform for this research project. This was selected primarily due to the ease of integration with Azure Machine Learning that we used for datamining and building the software applications required for this project (artefact). We needed to create Azure SQL subscriptions and then setup Azure SQL server to host the databases. These resources are all available on my personal subscription so I cannot provide the username and password as there are non-project related resources on my portal.

To create your Azure SQL DB then you need to go to the portal, create an account and then create the Azure SQL server instance and your DB's. The data can then be imported using SQL Server Management Studio and used the same way as any other SQL DB. The big advantage being is that your data is always available in the could with all the benefits of geo replication and backup etc.

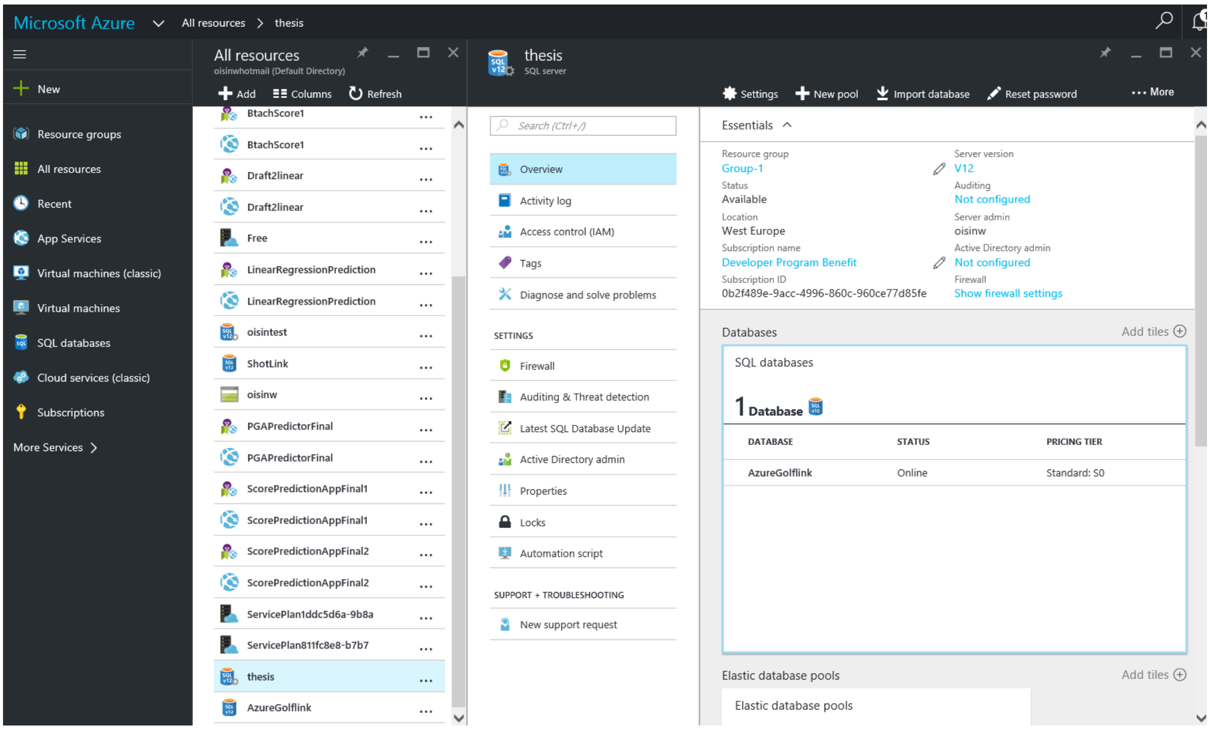


Figure 3: Working with the Azure Portal to create Azure SQL DB

Using R and R Code:

The 64-bit version of the R Language was used throughout this research project for exploratory analysis and statistical programming. This full code file was submitted with the project artefact on Moodle. This section will discuss the main work and include the appropriate snippets of code.

Development Environment and Libraries used:

Rstudio was used as the development environment and the following libraries were installed to run the required functions for this research:

- library(ggplot2)
- library(dplyr)
- library(chron) #install package
- library(corrplot)
- library(RODBC)
- library(normalp)
- library(relaimpo)
- library(car)
- library(rminer)
- library(psych)
- library(Hmisc)

Importing the data from Azure SQL:

Our data was stored in Azure SQL and in order to work with the data in R it had to be read in directly. The way I choose to implement this was using the RODBC's library 'odbcConnect' function and settling up Azure as an ODBS source on my windows client where I had R installed. Here is a sample query I used to create the graphs and models discussed in the project paper.

```
conn <- odbcConnect("Prediction", uid = "xxxx", pwd = "xxxx!")
odbcGetInfo(conn)
Eventdata <- sqlQuery(conn, "select  [Tournament Year], [Permanent Tournament Number], [Event Name],[MajorBin] , [Course Name], [Course Par] as  Course_Par, [Course Yardage],  MIN([Round 1 Score]) as Lowest_R1_Score, MIN([Total Strokes]) as Lowest_Total,[Average_R1_Score], [RND1_Leading_Score],(MIN([Round 1 Score] + [Round 2 Score]) - ([Course Par]*2)) as RND2_Leading_Score, (MIN([Round 1 Score] + [Round 2 Score] + [Round 3 Score] ) - ([Course Par]*3)) as RND3_Leading_Score, (MIN([Total Strokes]) - ([Course Par] * 4)) as Winning_Score, SUM([Birdies]) as Total_Birdies, SUM([Eagles]) as Total_Eagles, sum([Money]) as Total_Prize_Money

 from [dbo].[PredictScore]

 WHERE [Total Strokes] > 250

group by [Event Name], [Tournament Year], [Permanent Tournament Number], [Course Par], [Course Yardage],[Course Name], [MajorBin],[Average_R1_Score],[RND1_Leading_Score]

ORDER BY [Tournament Year] DESC;")
close(conn)
```

Once the data was available in R as a dataframe then this could be manipulated and used as per any other dataframe in R.

Linear Regression:

The R-code to create the simple linear regression models that are discussed in Section IV.A is included in the Rcode text file. Two linear models were created one to look at the leading Round 1 score and the other to look at the Average Round 1 Score:

```
mod <- lm(Eventdata$Winning_Score ~ Eventdata$RND1_Leading_Score)
mod <- lm(Eventdata$Winning_Score ~ Eventdata$Average_R1_Score)
```

R was used to review and test these models using the summary statistics and plotting the outputs. Note both the relationship was plotted and the residuals. The residuals plot was not included in the final document due to space restrictions. Here is the code for the first model:

```
#Scatterplot #1 for the paper:
################################
xtext <- "Rnd1. Leading Score"
ytext <- "Winning Score"
plot(Eventdata$Winning_Score ~ Eventdata$RND1_Leading_Score, pch = 21, bg = 2, xlab = xtext, ylab = ytext)
mod <- lm(Eventdata$Winning_Score ~ Eventdata$RND1_Leading_Score)
mod.res = resid(mod)
abline(mod, lwd = 2)
a <- round(coef(mod)[1], 2)
b <- round(coef(mod)[2], 2)
tx <- paste("y", " = ", a, " + ",b,'x ',  sep = "")
title(main = "Rnd1. Leading Score v Winning Score")
legend("topleft", bty="n", text.col = 4, cex = .9, legend=bquote(R^2==.(format(summary(mod)$r.squared, digits=4))))
legend("bottomright", bty="n", text.col = 4, cex = .8, legend=bquote(""*.(tx)))

########################################
#Plot the Residual's
plot(Eventdata$RND1_Leading_Score, mod.res, ylab="Residuals", xlab="R1 Average Score", main="Residuals Plot")
abline(0, 0)
```

```
#The Linear Model data to interpret
summary(mod)
cor.test(Eventdata$Winning_Score,Eventdata$RND1_Leading_Score)
cor(mod)
```

Correlation Analysis:

R was used for the correlation analysis discussed in section V.B of the paper. Three outputs from R for this were:

- Correlation Table – Table VI in the paper
- Correlogram – Fig 5. In the chart

Correlation Table:

To create the correlation table, the subset of the variables identified for correlation analysis were loaded into a data frame and the following function was used:

```
cor.test(finalset)
```

The output from this was a table with the Pearson's correlation coefficient and separately the p-values. I exported these results into Excel and formatted the table as per Table VI in the paper and used excel2latex to create the Latex code for the table. A lot of manual work involved here unfortunately.

Correlogram:

This is Fig 5. In the paper. This was run on the final seven variables to graphically display the correlation coefficients in order of strength so the important variables would stand out. The code to create this was as follows:

```
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(M, method="circle", col=col(200),
    type="upper",
    addCoef.col = "black", # Add coefficient of correlation
    tl.col="black", tl.srt=60,title="Correlation Matrix: Final Variables", #Text label color and rotation
    # Combine with significance
    p.mat = p.mat, sig.level = 0.01,  insig = 'blank',
    # hide correlation coefficient on the principal diagonal
    mar=c(0,1,2,0),diag=FALSE
```

The full code is included with the artefact the above is just the section that deals with creating the correlogram.

Multiple Linear Regression:

R was used to fit the multiple linear models discussed in section V.C in the paper. A number of models were fitted, mainly the initial model with all the features and the final model with just the significant features.

```
fullfit <- lm(testfeatures$Winning_Score ~ testfeatures$`Major` + testfeatures$Course_Par + testfeatures$`Course Yardage` +
testfeatures$Total_Prize_Money + testfeatures$Lowest_R1_Score + testfeatures$RND1_Leading_Score +
testfeatures$Average_R1_Score + testfeatures$RND2_Leading_Score + testfeatures$RND3_Leading_Score)
```

```
finalfit <- lm(testfeatures$Winning_Score ~ testfeatures$Course_Par + testfeatures$`Course Yardage` +
testfeatures$Total_Prize_Money + testfeatures$RND1_Leading_Score + testfeatures$Average_R1_Score)
```

The data from the summary of these models is what went into Tables VII and VIII in the paper. Once these models were fitted other functions were run to test for multicollinearity, perform stepwise regression

Descriptive Statistics used for discussion part of the paper:

```
summary(finalfit)
coefficients(finalfit)
residuals(finalfit) # residuals
anova(fit) # anova table
```

Feature Selection – stepwise regression

```
step(fit,direction = "backward") # se;ect a model and parameter based on AIC
```

#Testing for multicolinearity - A vif> 10 suggests collinearity.

```
vif(fit)
```

Relative Importance:

R was also used to produce the chart and data for discussion in section 5.D of the paper. The final 5 variables were put into a dataframe and the relampio function was used to create the statistical data and the chart which is fig 6. In the paper. The snippet of code to produce that is as follows:

```
calc.relimp(finalfit,type=c("lmg","last","first","pratt"),
        rela=TRUE)

# Bootstrap Measures of Relative Importance (1000 samples)
boot <- boot.relimp(finalfit, b = 1000, type = c("lmg",
                                "last", "first", "pratt"), rank = TRUE,
              diff = TRUE, rela = TRUE)
booteval.relimp(boot) # print result
plot(booteval.relimp(boot,sort=TRUE),main="Relative importance for 'Event Winning Score'",axes = FALSE,names.abbrev = 8) # plot result
```

The 'Predictfinal.R' file submitted along with the artefact has the full R code used for all sections of the analysis in the paper.

Microsoft Azure Machine Learning:

The Azure Machine Learning platform was used to evaluate and compare the machine learning algorithms and optimise the final two models to create web services for use as Azure Applications. The Azure applications are working pieces of software produced as the artefact for this research project. While many hours and weeks were spent modelling in Azure Learning Studio this manual will talk through the 3 main ones used for the paper.

Comparing the machine learning algorithms:

One main experiment was created to use apply each of the 5 algorithms on the data for this research. The first thing that was required for all experiments was to read the data in from Azure SQL. This was done via the 'Import Data' module. It was configured as per Figure 4 below. This essentially creates the connection to Azure SQL and then you write the custom SQL query for the data you need in the query box highlighted. The SQL query for this experiment is as follows:

```
select  [Tournament Year], [Permanent Tournament Number], [Event Name],[MajorBin] , [Course Name], [Course Par] as Course_Par, [Course Yardage],  MIN([Round 1 Score]) as Lowest_R1_Score, MIN([Total Strokes]) as Lowest_Total,[Average_R1_Score], [RND1_Leading_Score],(MIN([Round 1 Score] + [Round 2 Score]) - ([Course Par]*2)) as RND2_Leading_Score, (MIN([Round 1 Score] + [Round 2 Score] + [Round 3 Score] ) - ([Course Par]*3)) as RND3_Leading_Score, (MIN([Total Strokes]) - ([Course Par] * 4)) as Winning_Score, SUM([Birdies]) as Total_Birdies, SUM([Eagles]) as Total_Eagles, CAST(sum([Money]) AS FLOAT) as Total_Prize_Money

from [dbo].[PredictScore]
```

WHERE [Total Strokes] > 250

group by [Event Name], [Tournament Year], [Permanent Tournament Number], [Course Par], [Course Yardage],[CourseName], [MajorBin],[Average_R1_Score],[RND1_Leading_Score]
ORDER BY [Tournament Year] DESC;

Once the data is available in Azure ML studio the experiment can be built.

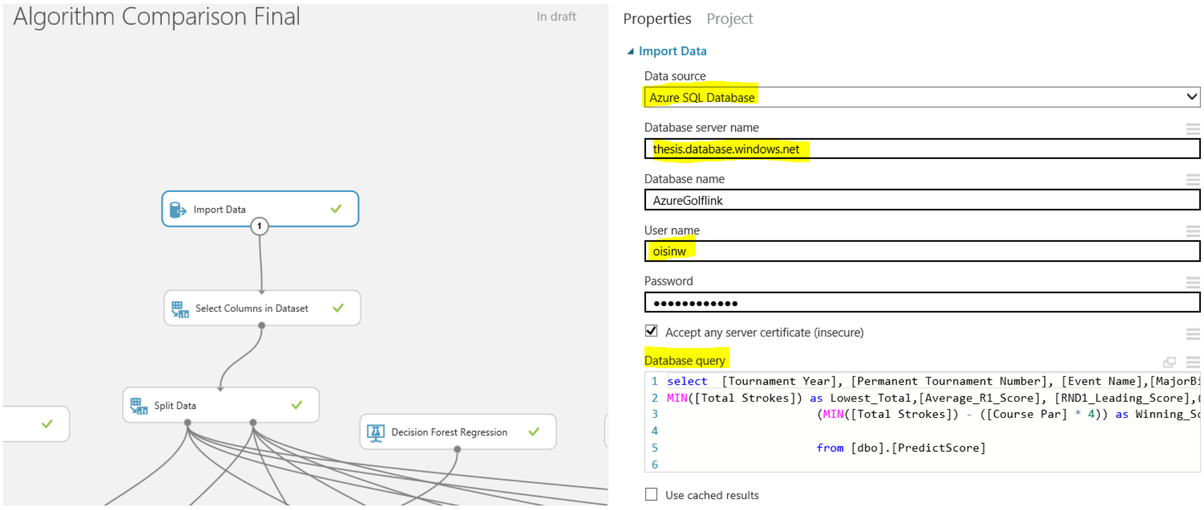


Figure 4: Configuring the Import module, connecting to Azure SQL and SQL query used to pull in the data

This next stage was to select only the features we were using to predict the winning score. These were selected via the statistical analysis in the paper. The select column module is configured as per figure 5. The features selected are the 5 predictor variables and the one dependent variable 'Winning Score'

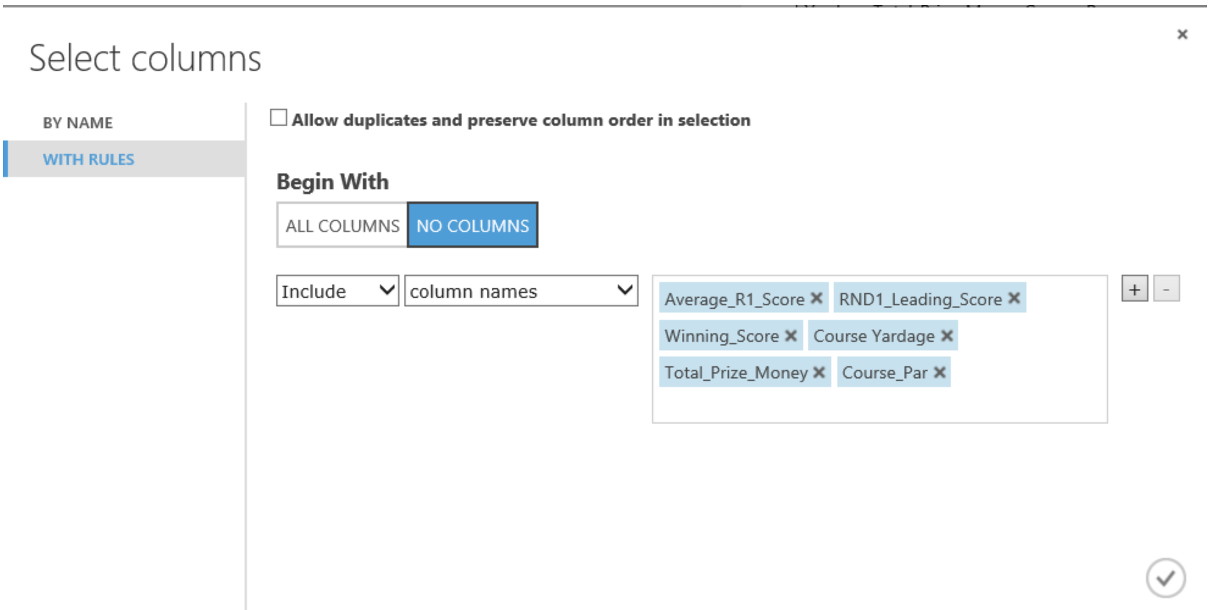


Figure 5The select column's module only the 5 predictor variables and the dependent variable were added

The last step in preparing the data was to split the data into the training and test set. The holdout and test process in described in the paper. The split data module is configured as per Figure 6.

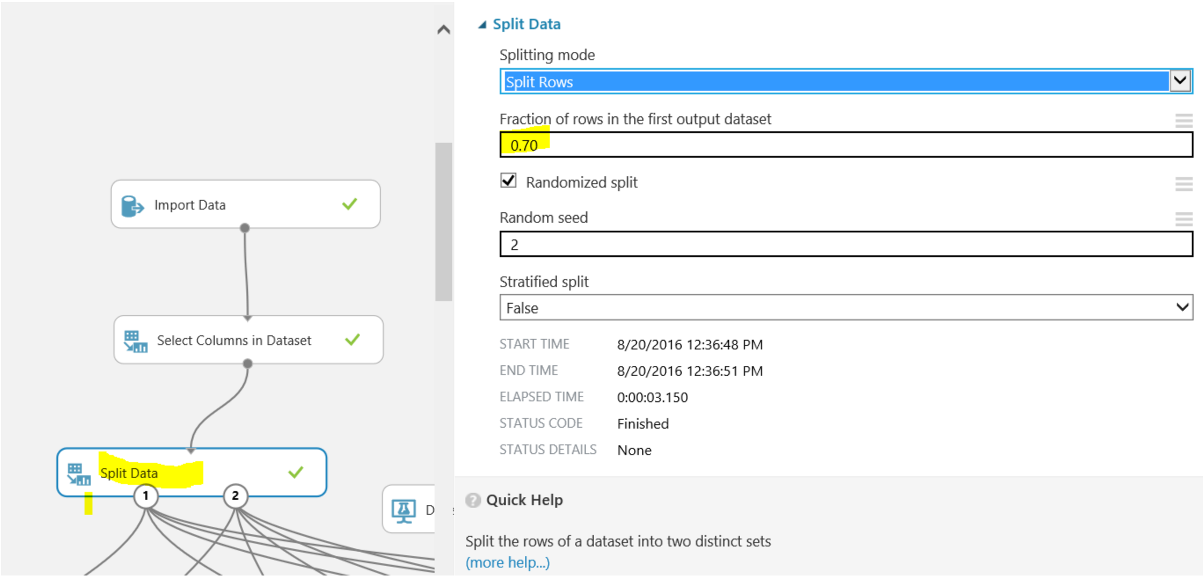


Figure 6: Splitting the data into 70% training and 30% for test

With the data imported and selected the next steps were to locate all the modules for the 5 algorithms namely 'Linear Regression', 'Bayesian Linear Regression', 'Decision Forest Regression' and 'Boosted Decision Tree' and 'Neural Network'. Each of these module had an associated 'Train', 'Score' and 'Evaluate Module'. The Train Module was configured to train on the Dependent variable winning Score (see Fig:7) and the Score module then stored the predictions while the evaluate module reported on the accuracy of the predictions.

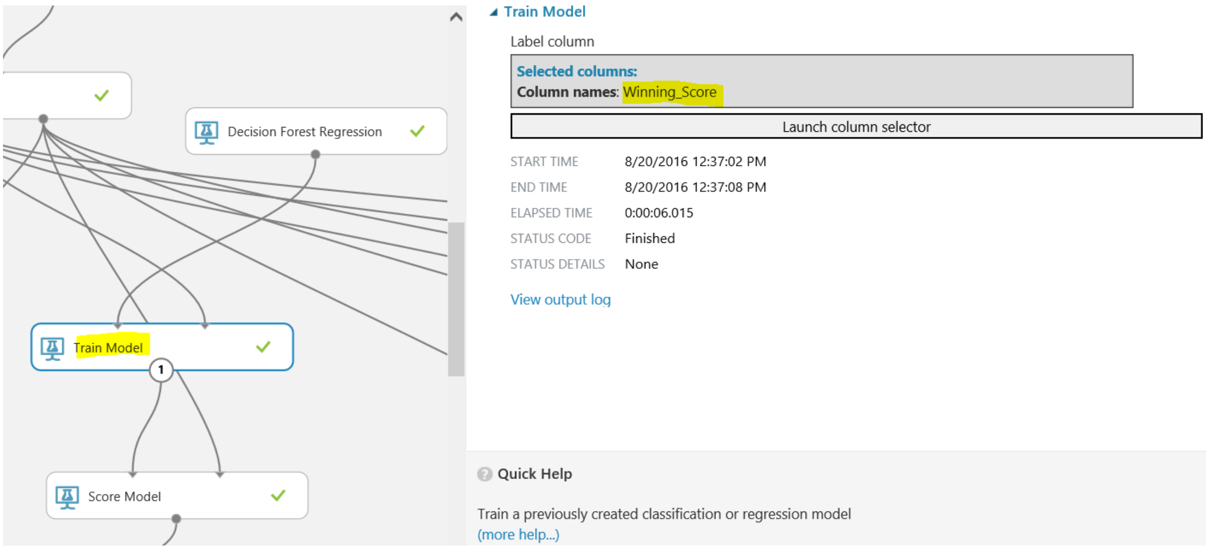


Figure 7: Configuring the Train Module to train on Winning Score

The full layout of the experiment can be seen in Figure 8. Once everything is linked up as per Fig 8 then select run and wait for all the boxes to turn green. To view the results then right click on the evaluate module to see the results as per Fig 9. I wrote custom R code to pull all the results together into one table for reporting the table in the paper.

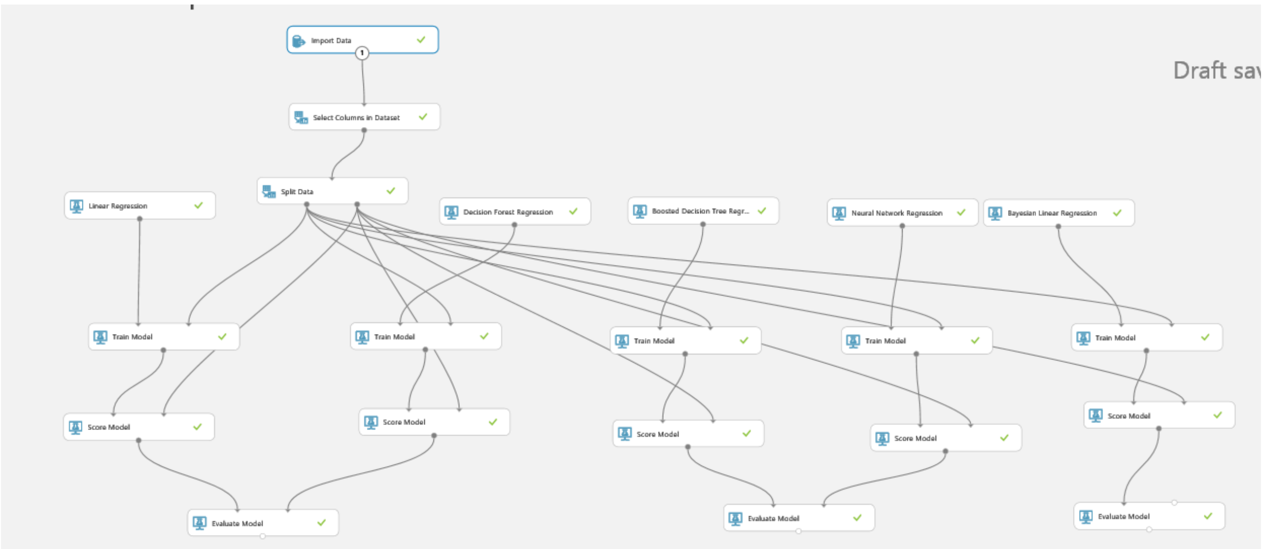


Figure 8: the full experiment with all the algorithms configured to Train, Score and Evaluate

Algorithm Comparison Final › Evaluate Model › Evaluation results

rows	columns
2	6

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Infinity	2.867861	3.686895	0.660294	0.420007	0.579993
433.780615	2.929297	3.698426	0.674438	0.422639	0.577361

Figure 9: Visualising the results from the evaluation module.

Optimising the Models:

The previous experiment walked through creating one to compare all algorithms. Once we selected the most accurate algorithms these were then built out as a separate experiment. These can be seen in Fig 10, and Fig 11. Below. Note the data was not split this time as it was trained on the full dataset. The 'Tune Hyperparameter' module added and what this does is configure the settings for each algorithm to the optimum settings for your data to ensure the most accurate model. Note the Cross-Validation module which is configured to run across 10 folds. These were created in the same way as outlined for the previous experiment.

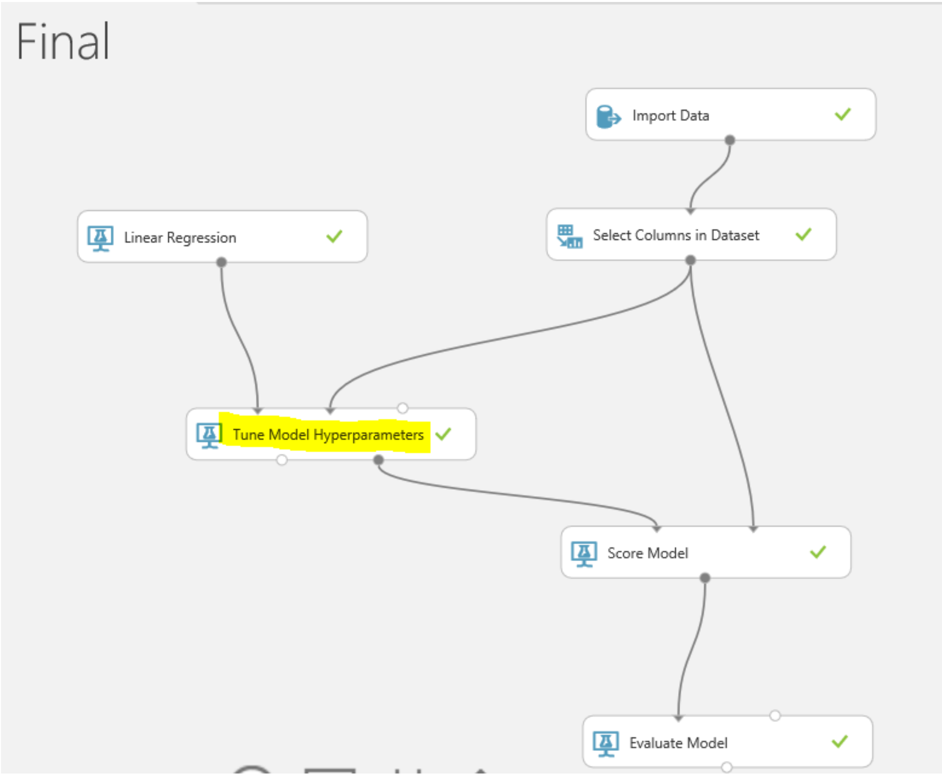


Figure 10 Sample Experiment with Tune Hyperparemters to tune the settings for the Linear regression algorithm to the optimum settings

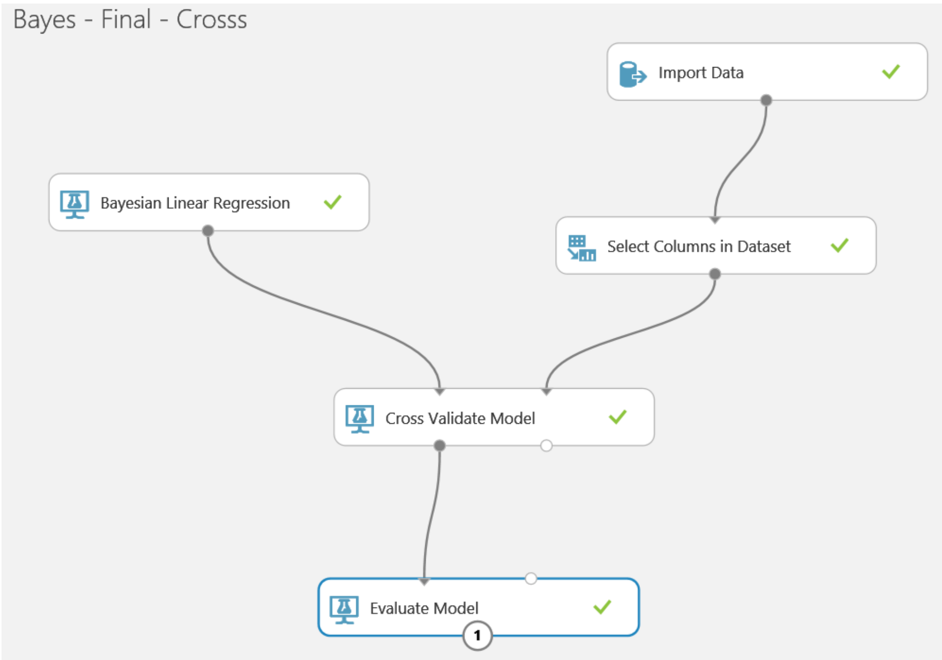


Figure 11: Final Experiment with 10-Fold Cross validation

Creating the Azure Applications:

Once the models were optimised these were deployed as predictive Web services directly from Azure. Figure 12. Shows one of the completed modules. Once these are deployed as REST API's we can then go to the Azure Portal and create the applications using the API key supplied from the Azure ML.

Once you've deployed the web service on Azure ML it can be managed in the Azure portal through the web service API. To do this you need to create a new Azure ML Request-Response Service Web App and configure it as per Fig 12. It will take a few minutes to create the application.

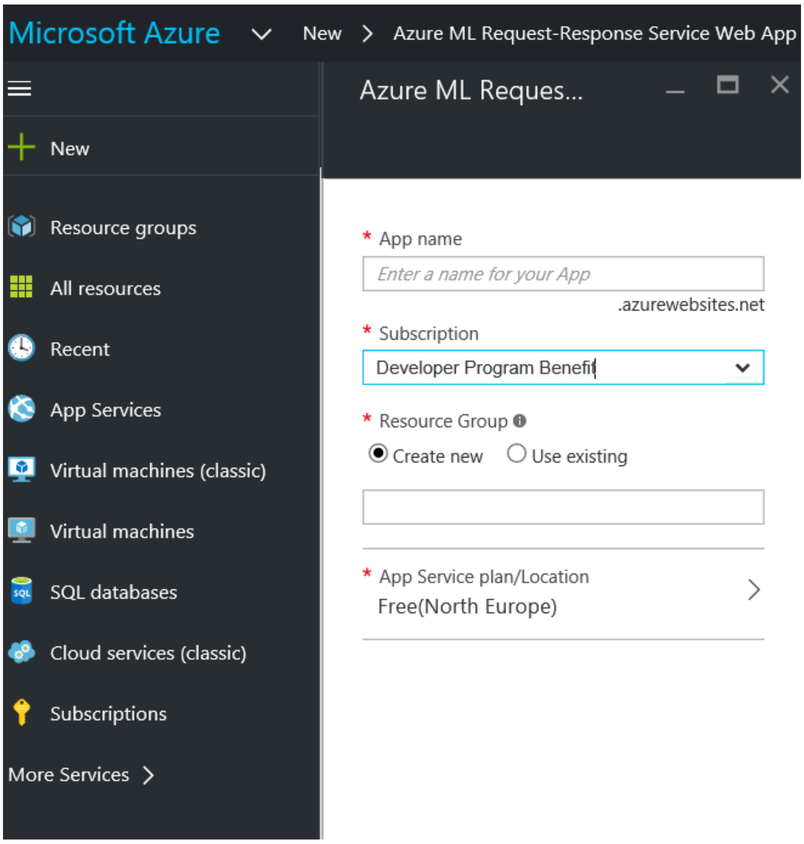


Figure 12: Creating the Azure ML Apps

Once it is created clicking the URL will ask you for the API key and once its entered brings you to the settings pages to create and customise your Application. Figure 13. Shows the web app being customised with proper names and explanatory text for the user. Submitting the changes launches your apps live for everyone to use. The deployed application is ready for use as per Figure 14.

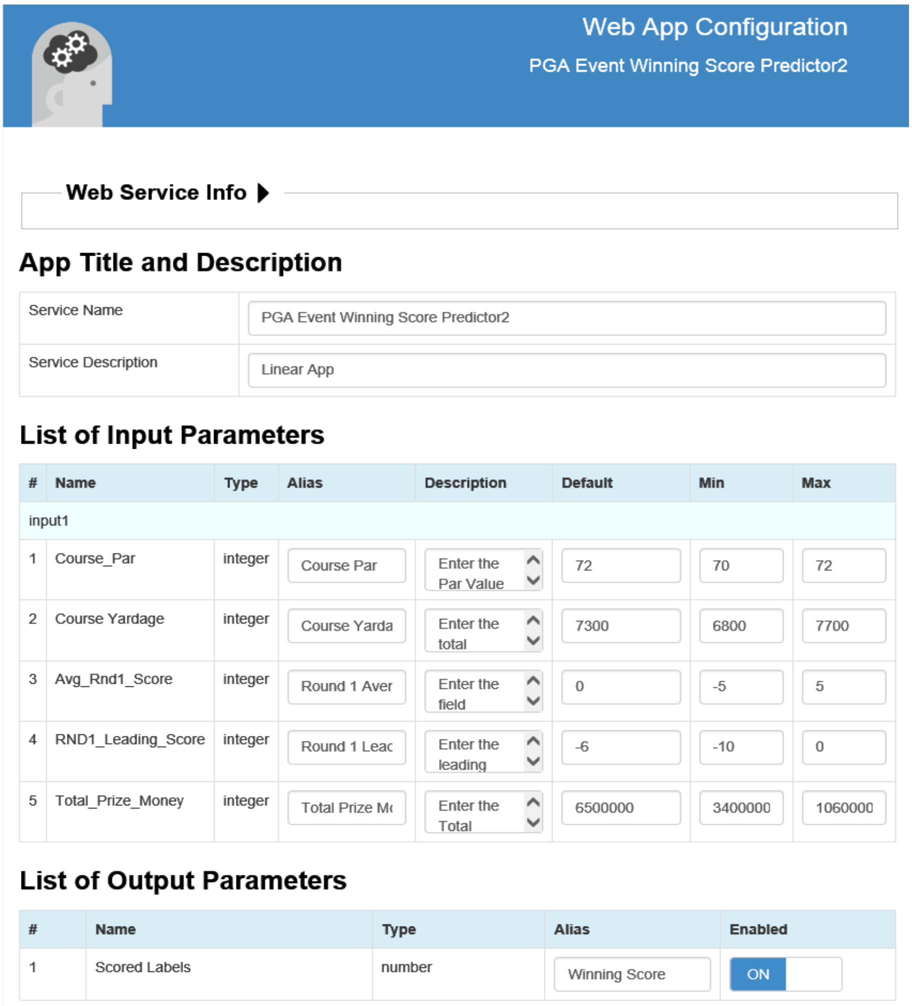


Figure 13: Customising the App on the main Azure Portal, allows you to change the UI, headings and explanatory text

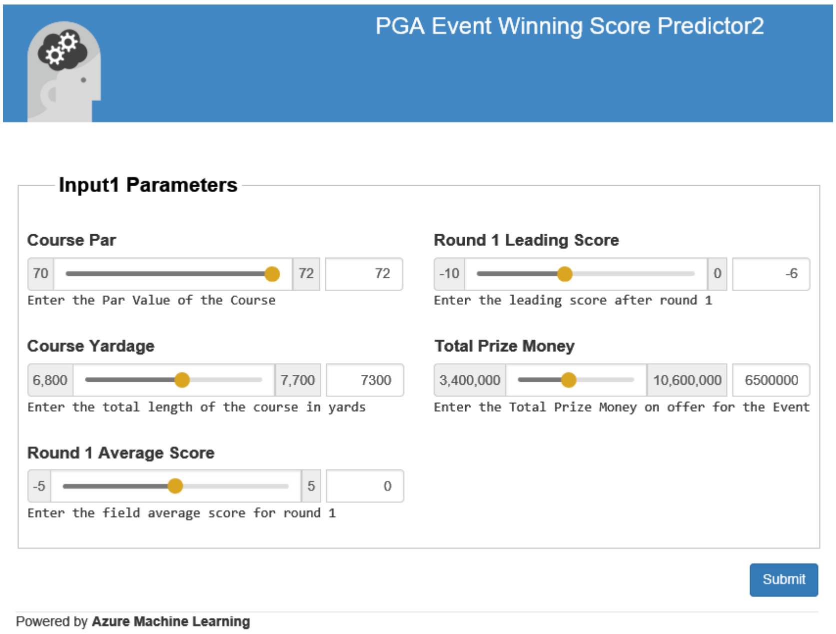


Figure 14: the final App deployed.

Running the final Azure Prediction Applications:

In order to run the final applications all, you need is a device connected to the internet. The URLS for the two web applications created for this research are:

The Linear Regression App: http://scorepredictionappfinal2.azurewebsites.net/

The Bayesian Regression App: http://scorepredictionappfinal1.azurewebsites.net/

Once you run the applications you will be asked to enter in the 5 input features required. These can be found on the PGA Tour Website www.pgatour.com . Figure 15 shows the tournament detail page that is available for each tournament with the 'Total Prizemoney', 'Course Par' and the 'Course Yardage'. The final two inputs 'Rnd1 Leading Score' and 'Rnd 1 Average Score' need to be manually sourced and calculated from the leaderboard after close of the first days play.

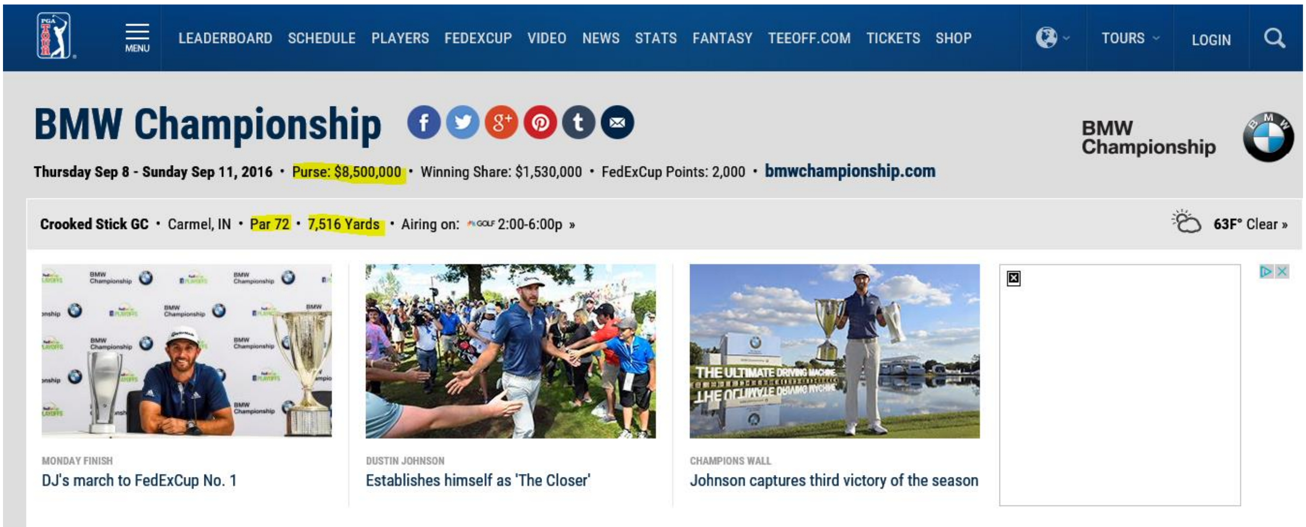


Figure 15: The prizemoney, course par and course yardage all available from the PGA TOUR website for the current tournament

You then run the app and fill out the fields as per Figure 16, hit submit and the predicted score is output for that event. Please go ahead and use them freely for the next tournament!

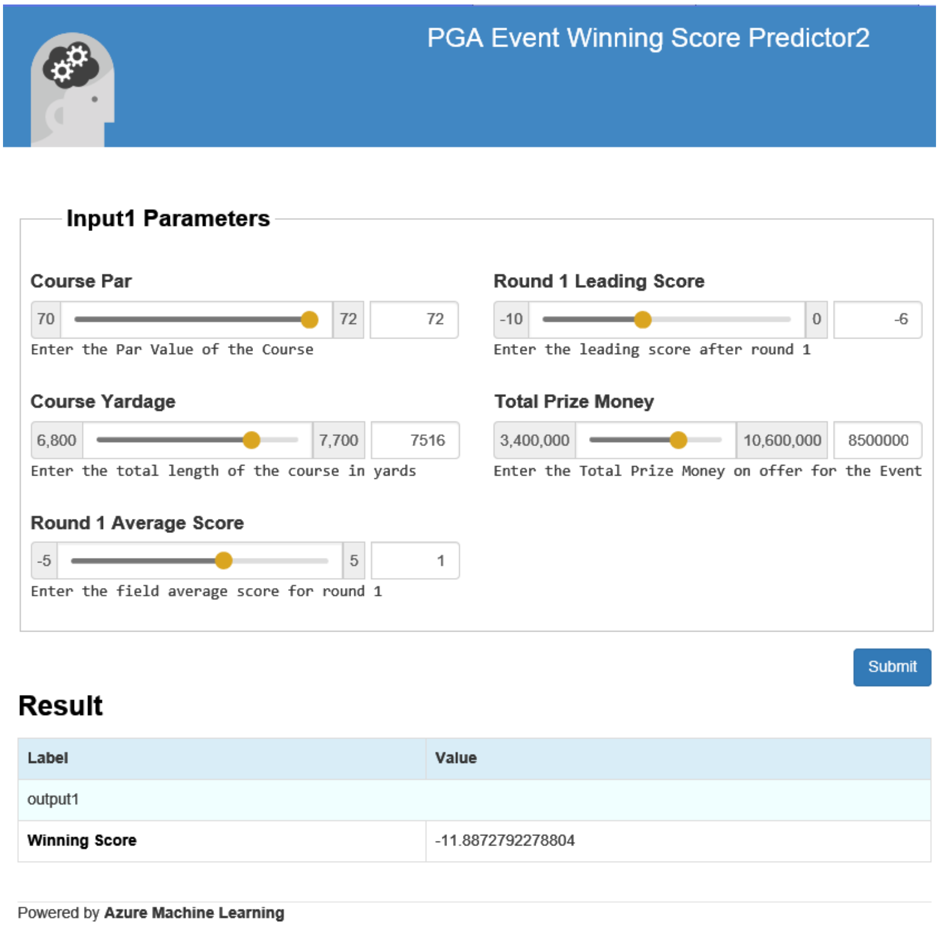


Figure 16: The deployed App in use showing the predicted Score for the current event

Conclusion:

This manual brought you through the main technologies used throughout this research project. It also went into detail on how the main technologies were configured. As the majority of the work is in the cloud these do not run offline or can't be packaged up in the same way as a client package can. However, all the work for this project will be kept on Azure for the next 6 months for inspection at any time. The artefacts produced are two working prediction applications that are live on the web for use by anyone. The information in this manual will enable you to re-create any of the experiments discussed in the paper, with the proviso that the data was released to me under NDA from the PGA tour so I trust that the college will treat the pre-processed data uploaded with the artefact in the strictest confidence.