**Report Structure**

**Business Opportunity**

For defined leagues - Give guidance on profitable odds as a service prior to each game

**Business Objective**

Identify breakeven probabilities per outcome per game

**Machine Learning Objective**

Develop a model that will consistently identify breakeven probabilities

Qualifier – Even if we fail to achieve the business objective - using the data we currently have, we will have built a framework that allows further development – add more data, create new features, develop better models.

**Understanding the Objective**

Odds and the Overround

Expected Value of a bet

**Machine Learning Project**

Re-statement of Machine Learning Objective

Data description – We will start with basic data – we need data that captures the game, and we need odds data to enable the financial calculations.

Data Sources and Data Wrangling

Exploratory Data Analysis

Match Results

Goals

Odds

Odds Variation

Shots

Shots on Target

Other Match Data

Expected Goals

League Statistics – Are Leagues Different? Test Goals assuming poisson, test goals/shots, SoT as binomial

<https://stats.stackexchange.com/questions/201247/equivalent-of-t-test-for-binomilal-poisson-variables>

Features

Across Season Features neglected

Models as Features

Poisson Regression

Odds as Implied Probabilities

Game Features

Data Work Flow

Metrics

Note on why Probability Calibration is required

Reliability Diagrams

Rank Probability Score

Expected Calibration Error

Field-Level Expected Calibration Error

Modeling

Dealing with imbalanced Classes

Feature Compression

Feature Selection

Model Comparison on validation Data

Model Results on held out test Data

**Business Results**

**Next Steps**

**Business Opportunity:**

**Soccer Sports Betting**

**Customer**

As more states legalize sports betting, the opportunities to provide services for bettors grows.

My business proposal is an web service that enables a bettor to identify a profitable bet on soccer games.

**Service Description**

The scope will be for 3 of the top European leagues:

* English Premier League – 20 teams playing 380 games per season
* German Bundesliga – 18 teams playing 306 games per season
* Italian Serie A – 20 teams playing 380 games per season

Each league is structured the same way. Every team plays every other team at home and away, and each game can has one of 4 outcomes:

* Home Team wins
* Draw
* Away Team Wins
* Game is abandoned, or some other unique event takes place – This is so rare that we will neglect it in this analysis

The web service will provide a list of upcoming games, along with the minimum odds required to make a profitable bet.

So, if for example on a Saturday, there are 20 games being played across these 3 leagues, 60 minimum odds recommendations will be made – 3 for each game – example shown below.



The “Profitable Odds” columns show **the minimum odds required** for a profitable bet.

**Customer Value**

The bettor will take these numbers and scan the Sportsbooks for odds that are higher than these values. If I consistently find and bet at odds higher than these, over the long run the bettor will be profitable

So if I want to bet on Liverpool winning their 21 January game against Manchester United, I would find a Sportsbook giving odds at 1.5 or higher, and place the bet.

**The random nature of Soccer Games**

There is a significant degree of randomness involved in soccer games. We can never really know with 100% confidence what the final score will be. This is very different to a classifying images of cats and dogs. For cats and dogs there is a solid ground truth for the class. A cat is a cat with 100% certainty. This can be determined before we run the image through the model. A random event is different, we can never get to the point where we are predicting a single outcome with 100% certainty, because the outcome itself is uncertain. We can never get to the point where we say …

However, if we do a thought experiment and imagine that a football game between the same 2 teams was played 100 times, we would reasonably assume that we would not get the same final score for all 100 games. In a study of reference book Luck is a major factor. In fact it could be argued that the inherent randomness in the game is what makes it so exciting, and hence so popular.

Because of the probabilistic nature of the outcome, we need to carefully consider the types of models we use to classify game outcomes.

**What makes a Profitable Bet?**

We can think of a football game like this. If this game were played a million times, how would the possible outcomes be proportioned? Would we expect to see the home team win 600,000 times, a draw 200,000 times, and an away win 200,000 times?

If we imagine that we have the ability to get this data, then we would be able to say that the probability distribution for the game results would be as follows:

Home Win Draw Away Win

0.6 0.2 0.2

The next component to the bet is the odds

Let’s say we go to 3 Sportsbooks and find the following odds for home wins – How do we know which bets are profitable?

**Expected Value**

We can assess profitability through an Expected value calculation

If we go back to the scenario where the game is played 1 million times, we can imagine we place the same bet 1,000,000 times and see how much better, or worse off we would be.

Note that when we place at $1 bet at 3.0 odds. If we win we get $3.0 back – our stake of $1, plus a profit of $2, so we are $1 better off. If we lose the bet, we lose our stake of $1, so we are $1 worse off

For home Win Bet at x odds

(600,000 x (3.0 – 1.0)) + (4000,000 x (-1)) = 600,000 - 400,000 = 200,000

So if we make this bet 1,000,000 times we expect to yield a profit of $200,000

However, we only get to see one game of the million, so we have to modify our calculation as follows:

(0.6 x (2.0 – 1.0)) + (0.4 x (-1)) = 0.6 – 0.4 = 0.2

The value of 0.2 is the Expected value. It is the probability distribution for a set of outcomes multiplied by the profit or loss associated with that outcome

EV = probability of event occurring x potential profit/loss due to the event + probability of event not occurring x potential profit/loss if the event does not occur

We can apply this to the odds from our bookmakers, assuming we are betting a unit

EV = p(outcome) x potential profit + p(other outcomes) x potential loss

EV = p(outcome occurs) x (odds -1 ) + (outcome does not occur

are betting on 100,000 games we blah blah blah

**Calculating Expected Value**

EV = probability of event occurring x winnings – probability of event not occurring x loss

Let’s say we find the following odds:

Home Win Draw Away win

Odds xx xx xx

Probability of event the model predicts











We will only bet if we find positive expected value

**Comments:**

* We have made a significant assumption – the model probability is good enough to enable us to make the EV calculation with confidence
* Higher Odds may convert a negative EV into a positive EV, so we should look for the best (highest) possible odds
* Our model makes 3 probability predictions. If one probability goes up, then either one or both of the other probabilities goes down. We may have 2 positive EV bets on a single game – Theoretically, we should place both bets. Practically, … logistics issues in getting $ on

We could place every positive EV bet. Alternatively, we could use a bet selection strategy. For example, if we have 2 positive EV bets on a single game, we could choose the highest EV, and make a single bet. We could modify select the 3 highest EV bets out of all the games being played that day. We could have a model that seems to perform better on estimating Home Win probabilities, and so only place Home Win bets. We may decide only to take low odds bets, because high odds bets are too volatile.

|  |
| --- |
| Select all positive EVS |
| Select highest EV per game |
| Select lowest positive EV per game |
| Select bets with Evs above a threshold |
| Select positive EV with odds below a threshold |
| Select positive EV bets where max probability difference is less than threshold |
| Select where maximum absolute difference between implied and model probabilities is below a threshold |
| Select where positive EV, and maximum model to bookie odds difference odds is lower than a reasonable (empirically determined) threshold |

**Scoring the Model**

The objective of the project is to identify profitable betting opportunities. A profitable betting opportunity is captured if we are good at estimating its’ Expected Value. Expected Value is determined by 2 inputs:

* The probabilities output by the model
* The odds given by the bookmaker

We have to remember that our model is producing probabilities of an event, but the event is itself somewhat random.

We cannot control the odds given by the bookmaker, but we can try and develop a model that is “good” at predicting the outcome probabilities. Note that this is different to accurately classifying outcomes. How do we assess the probability output from the model?

The answer lies in a 2011 paper “Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models” This paper shows that soccer results can be thought of as an ordinal ranking. Home win is first, followed by a draw, followed by an away win, and that tool called the Rank Probability Score is a better way to assess predictions.

What is the Rank Probability Score?

This measures how good a probability forecast is at classifying an observed outcome. A perfect score is 0, the worst possible score is 1

Formula goes here

**Consider Extreme Predictions, and the impact on RPS**

It is worthwhile reviewing some examples of Rank Probability Scores for the 3 possible match outcomes – home win, draw, and away win



This table is enlightening. It is comprised of extreme prediction probabilities where we predict an outcome with absolute certainty – Our forecast probabilities are either 0 or 1

* For RPS a lower score is better
* We score an RPS of 0, when we predict an outcome with a probability of 1
* The best possible RPS for all outcomes – home win, draw, away win is 0
* The worst possible RPS for an observed Home or Away Win is 1.0
* The worst possible RPS for an observed Draw is 0.5 – this is the impact of a draw being the middle of the 3 outcomes

**Consider Baseline Frequency Predictions, and the impact on RPS**

European soccer has a significant home field advantage. This varies over time, but a home win is far more likely than either a draw, or an away win. In fact a home win is about twice as likely.

We can plug some rough baseline frequencies into our RPS calculation and determine and review the results.



Some comments on this table:

* The RPS for each baseline frequency is different depending on the actual outcome
* There is a significant difference between the best (lowest) Draw, and the highest (worst) Away Win
* This means that just using baseline frequency predictions, we do far better at predicting draws, than we do at predicting either home Wins, or away Wins.

**Graphical Summary of Tables**

**Proposal for Model Scoring**

We need to devise a method of rating models that takes into account the quirks of the RPS, and the data.

Each model will produce a set of 3 probability scores between 0 and 1 for all possible outcomes of each game. These 3 probabilities are converted to a single RPS between 0 and 1. A model will produce a distribution of RPS values – one for each game it is predicting. This distribution may not be symmetric, and therefore we may not want to use the mean to capture the distribution location. The median which splits the sample at the 50% point may be a better measure of distribution location. However, because the baseline frequency RPS is different for each match outcome, we will compare the RPS median to the baseline frequency RPS. If the distribution median is lower, then we will assume the model is better than if the median was greater than the baseline frequency.

However, each outcome has a different baseline RPS. Therefore, we will subtract the model median RPS for each game outcome from the baseline RPS, and normalize by dividing by the baseline RPS

This is more easily seen on a chart

We can average these 3 numbers together to get an overall model score.

To get a score for a model

1. Estimate the baseline frequency RPS for each game outcome – baseFreqRPS\_hwin, baseFreqRPS\_draw, baseFreqRPS\_awin
2. Predict the probabilities of each outcome for all games in the data set
3. From these probabilities, calculate the RPS for each game
4. For each observed outcome – Home Win, Draw, away Win – Use the game RPS scores to calculate the Median RPS of the distribution – ModelMedRPS\_hwin, ModelMedRPS\_draw, ModelMEDRPS\_awin
5. For each observed outcome, subtract the model median RPS score from the baseline frequency RPS, and divide by the baseline frequency RPS to get 3 values
6. Take the mean of the 3 values to get an overall score for the model

**Data**

|  |  |  |
| --- | --- | --- |
| Strategies |  |  |
|  | Select all positive EVS | |
|  | Select highest EV per game | |
|  | Select lowest positive EV per game | |
|  | Select bets with Evs above a threshold | |
|  | Select positive EV with odds below a threshold | |
|  | Select positive EV bets where max probability difference is less than threshold | |
|  | Select where maximum absolute difference between implied and model probabilities is below a threshold | |
|  | Select where positive EV, and maximum model to bookie odds difference odds is lower than a reasonable (empirically determined) threshold | |

**EDA**

Goals scored as a Poisson Distribution

Home Field Advantage & baseline Result frequencies

Odds as Implied probabilities

**Modeling**

Poisson Regression

Goals scored as a Poisson Distribution

Poisson Regression Model

Negative Binomial Generalized Linear Model

Logistic regression

Ordinal Logistic Regression

A note on Expected Goals

Mean Odds Model

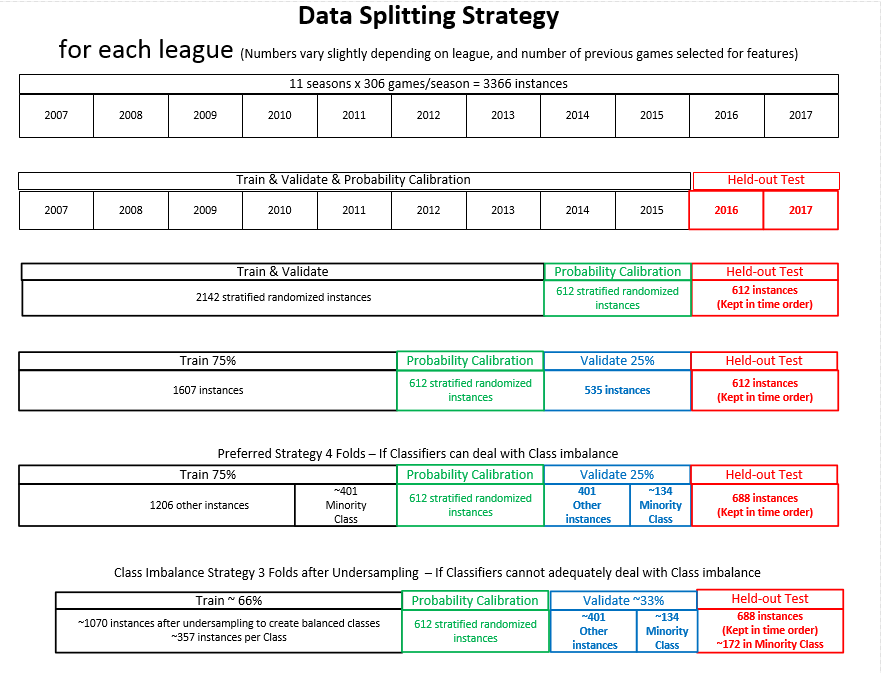
Model results

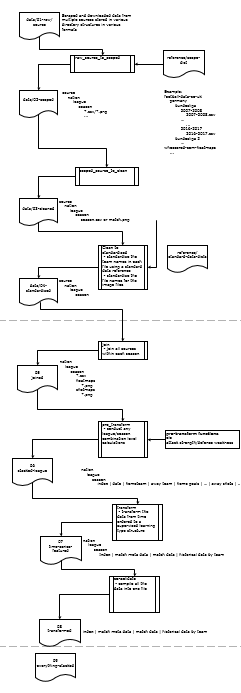
RPS Distributions as boxplots

**Betting**

Calculating minimum profitable odds from mean odds

Conclusion









**Notes**

binomial distribution

**Betting Performance Table**

Confusion matrix for placed bets vs actual outcomes with the boxes filled with $ values

Assessing Betting Performance

Single sided t test against 0

Overall Process

Data Pipeline -> Modeling Pipeline -> Gambling Strategy

Data Pipeline

The data was collected from football-data.co.uk, and indatabet.com

A pipeline was written to carry out the following steps

Process Flow Diagram goes here as per existing code

Exploratory data Analysis

Odds as Implied Probabilities, Poisson Distribution of Goals

Rank Probability Skill Score - Measure

Modeling Process

Add Features, Transform Data, Select Features, Tune parameters,

Results

Which features are important?

Baseline Model

Which is the

My Models

Logistic Regression

XGBoost

Conclusion

Verdict

Next Steps

How to implement – system diagram

A guide to calibration plots in python

<https://changhsinlee.com/python-calibration-plot/>

Scikit correct way to calibrate classifiers with CalibratedClassifierCV

<https://stats.stackexchange.com/questions/263393/scikit-correct-way-to-calibrate-classifiers-with-calibratedclassifiercv>

How and When to use a calibrated classification model with scikit-learn

<https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/>

An introduction to Probability calibration

<https://blog.cambridgespark.com/probability-calibration-c7252ac123f>

Classification probability calibration

<https://www.kaggle.com/residentmario/notes-on-classification-probability-calibration/>

Probability Calibration for Imbalanced Dataset

<https://towardsdatascience.com/probability-calibration-for-imbalanced-dataset-64af3730eaab>

We will simulate betting as close as we can get to the real world. We will assume we have a system that scans odds, and can return the mean odds. We will use this to calculate the Expected Value of the bet. However, when we place the bet, we will look for the maximum odds, and place the bet at these odds.

This may not perfectly represent what we can do in reality because at the time of placing the bet the mean, and maximum odds may be different to the values in the data. However, we will consider this strategy as reasonable for the purposes of model analysis.

In summary:

* Calculate Expected value using Mean Odds
* Calculate bet return using maximum odds