Métodos de Aprendizado de Máquinas aplicado à predição de jogos de futebol

Felipe Augusto Pereira Fernandes FELIPEAPFERNANDES@GMAIL.COM
CEFETMG

Abstract

Futebol é considerado o esporte mais popular do mundo (de Football Association et al., 2007). Por isto, é muito comum vermos pessoas discutindo em reuniões familiares e mesas de bar sobre quem vai ganhar uma partida. Analisando os resultados dos jogos das equipes é possível estabelecer um histórico de desempenho durante a temporada. Este trabalho estuda algoritmos de aprendizado de máquina para prever o resultado de uma partida de futebol: vitória do mandante, empate ou vitória do visitante. A partir de dados básicos de desempenhos e com entradas que tentam capturar as influências da dinâmica de uma partida de futebol, este trabalho tem o objetivo de estudar o desempenho dos algoritmos de aprendizado de máquina aplicados a predição de jogos de futebol, utilizando uma base de dados pública, (kaggle, 2016), que contêm informações de jogos de 8 países.

Keywords: Soccer match prediction, Pattern Classification, Machine Learning, Football.

1. Introdução

Futebol é o esporte mais popular do mundo e no Brasil não é diferente. Futebol é um assunto sempre presente em mesas de bares, reuniões familiares e grupos de amigos. Qual time vai ganhar a partida? Quem vai vencer o campeonato? Qual time é melhor? são indagações sempre presentes nas conversas. Justamente por isso, veículos de comunicação como jornais e programas de televisão sempre possuem especialistas para discutir e fazer projeções, empirícas, sobre quem serão os vencedores dos jogos da rodada do campeonato.

Determinar o resultado de uma partida não é uma tarefa simples, pois o futebol é um esporte coletivo sujeito a inúmeras variáveis, como padrão de um time, esquema tático, falhas individuais de jogadores, capacidade de execução e habilidade dos jogadores, condição do gramado, motivação da equipe, montagem da equipe, árbitros etc.

Um estudo, apresentado em *The numbers game: why everything you know about football is wrong* (Anderson & Sally, 2013) e feito com mais de dez mil jogos, mostra que a média de gols por partida vem mudando ao longo dos anos. Atualmente, essa média é de 2.6 gols por jogo. A Figura 1 mostra o comportamento desta média nas temporadas do campeonato inglês da primeira divisão desde 1940 a 2014.

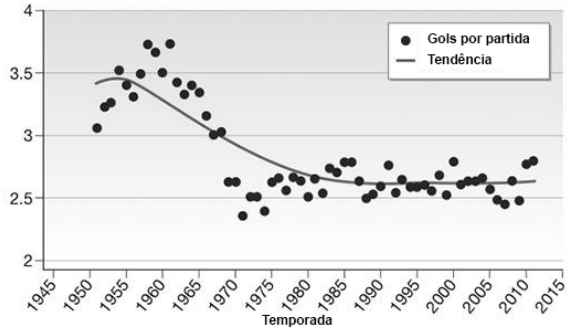

Figura 1. Média de gols em uma partida ao longo dos anos.

Ao analisar os resultados dos jogos das equipes durante uma temporada é possível estabelecer um histórico de desempenho, como quantidade de gols marcado, gols sofrido, quantidade de vitórias, derrotas, empates e etc. A partir destes valores é possível perceber que diferentes times apresentam valores próximos em momentos distintos do campeonato, quando os valores comparados são divididos pela quantidade de jogos já realizados. Desta forma, é possível estabelecer o padrão de um time em um determinado jogo em uma determinada rodada em um campeonato.

Estatísticas de jogos, times e jogadores são utilizadas no auxílio a determinação de resultados de jogos de futebol. Hoje existem sites especializados em estatísticas, que vem sendo amplamente divulgados em programas de televisão e em transmissão de jogos. Analisando os resultados dos jogos das equipes é possível estabelecer um histórico de desempenho durante a temporada. Quando os números das

Machine Learning Class – CEFET-MG, Belo Horizonte, Brasil, 2016.

equipes são analisados percebe-se que podemos descrever uma partida de futebol com esses números de ambas as equipes.

Este artigo propõe o estudo de modelos de aprendizado de máquina para a predição de jogos de futebol. Estudar seus impactos e performance em cima da base de dados pública. Além disso, este artigo visa a consolidação dos estudos na disciplina de aprendizado de máquina, cursada no mestrado em modelagem e matemática computacional do CEFETMG.

2. Trabalhos relacionados

Apesar da grande popularidade do esporte, existem poucos trabalhos científicos dedicados à predição de jogos de futebol (Aslan & Inceoglu, 2007) . Em Cheng et al.(Cheng et al., 2003), os autores demonstram o bom desempenho da utilização de Redes Neurais Artificiais (RNAs) para a predição de resultados de jogos de futebol. No trabalho, as RNAs implementadas são comparadas a métodos estatísticos, apresentando taxas de acerto superiores. Cheng et al. (Cheng et al., 2003) propõem o uso de RNAs para a predição de jogos de futebol. A Figura 2 exemplifica o modelo proposto. O modelo baseou-se em uma abordagem hibrida, criando um sistema que primeiramente avalia o nível dos times, separando os jogos em 3 conjuntos: quando o time mandante é melhor, times de níveis iguais e o time visitante é melhor. Para isto foi utilizado a diferença da média de pontos e a diferença da média de gols feitos pelas equipes. Para o cálculo da média de pontos, calcula-se os pontos dos times na competição até uma determinada rodada. Para cada vitória soma-se três pontos, para cada empate um ponto e zero pontos para a derrota. Ao final, divide-se a quantidade de pontos somados pela quantidade de jogos disputados pelo time. A média de gols feitos pelas equipes é calculada somando os gols feitos até determinada rodada e então divide-se esta soma pela quantidade de jogos disputados.Cada entrada da rede passa por este mecanismo seletor que define a qual RNA aquele jogo pertence. O modelo de entrada utilizado pelos autores é apresentado na Figura 3.

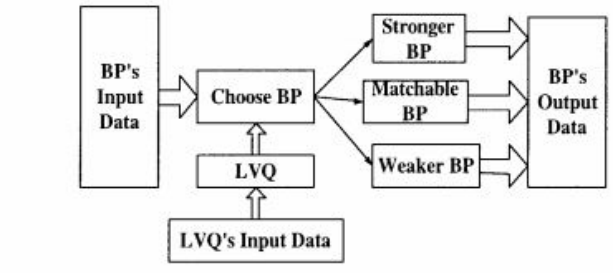

Figura 2. Architecture model proposed by Cheng et al. (Cheng et al., 2003)

Para o treinamento eles utilizaram os dados do campeonato italiano da tempo- rada 2001-2002, com 34 rodadas. Inicialmente foram utilizados os dados da rodada 7 até a rodada 17 para treinamento. O teste foi feito com os jogos da rodada 18 até a 34. Entretanto o teste foi feito rodada a rodada, somando 17 taxas de acertos, uma para cada rodada. Ou seja, para o teste da rodada 20, os dados dos jogos até a rodada 19 foram utilizados para o treinamento da rede, e a taxa de acerto foi coletada, utilizando sempre 9 jogos para realizar o teste. Os jogos das 6 primeiras rodadas não foram utilizados no processo de treinamento da rede nem de teste já que os autores acreditam que os primeiros jogos estão muito mais sujeitos às aleatoriedades do futebol.

A saída dessa rede híbrida seria a predição do resultado do jogo como vitória mandante, empate ou vitória do visitante. Este modelo proposto atingiu um taxa de acerto de 52.29% sendo melhor que os métodos de previsão estatísticos existentes, tais como Elo Rating (Elo, 1978). Entretanto Cheng et al. (Cheng et al., 2003) apontam o fato de que as últimas seis rodadas apresentaram uma taxa de acerto de 44.77%, justificando esta porcentagem ao número limitado de jogos para o treinamento e a fatores acidentais a que o futebol está sujeito.

Aslan e Inceoglu (Aslan & Inceoglu, 2007) fizeram um comparativo de modelos de previsão de resultados de jogos de futebol. Ao fim propuseram um novo modelo de rede, baseando somente no desempenho de cada time como mandante e visitante. Para isso, eles simplesmente incrementam uma variável para cada vitória, decrementam para cada derrota e em caso de empate não alteram o valor da variável. Foram propostos então dois modelos de entradas para redes Learning Vector Quantization (LVQ): uma somente com os dados de desempenho como mandante para o time mandante e os dados de desempenho como visitante para o time visitante; e outra com os dados de mandante e visitante para ambos os times. Os processos de treinamento e teste foram idênticos aos propostos por Cheng et al. (Cheng et al., 2003). A rede com menos entradas teve 53.25% de taxa de acerto contra 51.29% do outro modelo. Os autores concluíram que a modelagem dos dados de entrada da RNA é um problema crítico.

3. Base de dados

Para este estudo uma base de dados pública foi utilizada. Ela está disponível no *site* kaggle (kaggle, 2016). Nesta base de dados tem mais de 25000 partidas, onde tem informação de mais de 10000 jogadores de 11 países com seus principais campeonatos. Esta base de dados também tem informação sobre as escalações e posições dos jogadores(coordenadas X e Y) e valores dos prêmios pagos pelas casas de apostas para vitória do mandante, empate e vitória do visitante.

$$\begin{bmatrix} Team(A) \\ Team(*) \end{bmatrix} = \begin{bmatrix} win-ratio(A) \\ draw-ratio(A) \\ lose-ratio(A) \\ average-goal(A) \\ average-lose(A) \\ morale(A) \\ hom\,e-away(A) \\ win-ratio(*) \\ draw-ratio(*) \\ lose-ratio(*) \\ average-goal(*) \\ average-lose(*) \\ morale(*) \\ hom\,e-away(*) \end{bmatrix}$$

Figura 3. Modelo de entrada proposto por Cheng et al. (Cheng et al., 2003)

Esses dados estão representados em sete tabelas: *Country*, *League*, *Match*, *Player*, *Player Attribute*, *Team* e *Team_Attributes*.

4. Modelagem

Como o futebol é um esporte coletivo, suscetível a várias aleatoriedades, no qual fatores físicos, técnicos e psicológicos influenciam no desempenho de uma equipe, mensurar a qualidade de uma equipe, transformando essa qualidade em um número, é algo altamente subjetivo e complicado.

Em um jogo de futebol pode-se dividir uma equipe em três setores: defesa, meio de campo e ataque. Em geral, os especialistas afirmam que times com defesa e meio de campo fortes tendem a sofrer menos gols, enquanto times com ataque e meio de campo fortes tendem a fazer muitos gols. Portanto, indicadores de qualidade da defesa, meio de campo e ataque dos times parecem ser entradas interessantes para a modelagem desse problema. Baseado em especialistas, discretiza-se a qualidade das equipes, conseguindo fazer uma comparação da qualidade das mesmas.

Para obter esses valores, foram utilizados os indicadores de qualidade de defesa, meio de campo e ataque da franquia de jogos da EA Sports (Electronic Arts Sports), o FIFA. Nele, os jogadores são avaliados de 0 a 100 em diversos atributos físicos, técnicos e psicológicos. O índice do time é calculado como a média dos atributos de todos os jogadores da equipe. Desta forma, o índice de defesa do time é calculado pela média dos índices defensivos de todos os jogadores da equipe, o índice de meio de campo do time pela média dos atributos de meio de campo e o índice de ataque do time pela média dos atributos de ataque dos jogadores.

É sabido que o mando de campo tem influência no desempenho de um time, pois jogar ao lado de sua torcida ou sob a pressão da torcida adversária exerce um efeito psicológico no time. Portanto para a predição de um jogo é interessante avaliar o desempenho do time mandante nos jogos em que ele atuou em casa e o desempenho do time visitante nos jogos em que atuou fora de casa.

No futebol suspensões, lesões e transações de jogadores entre equipes é algo muito comum durante um campeonato. A mudança de um elenco durante um campeonato pode afetar o desempenho da equipe. Para que a rede seja capaz de captar essas variações é interessante levar em consideração o desempenho da equipe nos últimos jogos.

Em relação ao impacto da suspensão, lesões e transferência de jogadores, avaliar a performance de um time desde o ínicio da temporada não parece ser a melhor maneira.

Se colocarmos esses indicadores estatisticos de desempenho e analizarmos a performance de uma forma mais local, nos últimos 5 jogos, parece ser mais interessante, visto que esses indicadores mudam de valor quando aplicamos essa janela de tempo no desempenho.

Analizando localmente parece ser mais fácil predizer um próximo comportamento. A figura 4 demonstra os gols por jogo de uma forma global, desde o ínicio da temporada. Já a figura 5 demonstra o indicador gols a favor por jogo de uma forma local, considerando somente os últimos 5 jogos. As figuras esclarecem essa diferença nos valores dos índices quando aplicamos essa janela de tempo no desempenho.

Figura 4. Gols marcado por Liverpool na temporada 2014-2015.

No campo date, foi calculado a diferença em dias do jogo em questão para o ultimo jogo de ambas as equipes, criando assim uma data para o mandante e uma para o visitante.

Portanto estas ideias expostas nesta seção farão parte da descrição de uma partida de futebol.

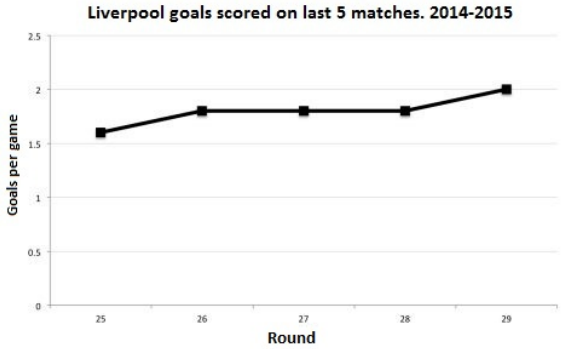

Figura 5. Gols marcados por jogo nos últimos 5 jogos do Liverpool na temporada 2014-2015.

5. Metodologia

Para este trabalho irei aplicar o algoritmo de regressão logistica, *KNeighbors*, *Random Forest* e *Gradient Boosting* ao dado extraído da base de dados.

Jogos dos campeonatos inglês, alemão, francês, espanhol, italiano e português foram utilizados, da temporada 2010-2011 até 2015-2016.

Após aplicar os dados puros da tabela *Match* irei aplicar um *Feature Engineering* para extrair mais features dos dados presentes nas outras tabelas.

Primeiramente usaremos somente os dados disponiveís na tabela *Matches* para fazermos a descrição de uma partida de futebol. Selecionei todos os jogos onde eu tivesse informação de todos os jogadores da partida e que tivesse dado a respeito do prêmio pago pela casa de aposta Bet365.

Após essa primeira filtragem, foram excluidos os dados referentes aos incidentes da partida, pois poucos jogos continham alguma informação, em formato XML, nestes campos.

Depois uma transformação de variáveis categoricas - *date*, *season*- para ordinal foram feitas.

Por fim, inclui uma coluna referente a classe do jogo: 'H' para vitória do mandante, 'D' para empate e 'A' para vitória do visitante.

Esses dados são compostos pelas seguintes caracteristicas:

'country_id', 'league_id', 'season', 'stage', 'date', 'match_api_id', 'home_team_api_id', 'away_team_api_id', 'home_player_X1', 'home_player_X2', 'home_player_X3', 'home_player_X4', 'home_player_X5', 'home_player_X6', 'home_player_X7', 'home_player_X8', 'home_player_X9', 'home_player_X10', 'home_player_X11', 'away_player_X1', 'away_player_X2', 'away_player_X3', 'away_player_X4', 'away_player_X5', 'away_player_X6', 'away_player_X7', 'away_player_X8', 'away_player_X9', 'away_player_X10', 'away_player_X11', 'home_player_Y1', 'home_player_Y2', 'home_player_Y3', 'home_player_Y4', 'home_player_Y5', 'home_player_Y6', 'home_player_Y7', 'home_player_Y8', 'home_player_Y9', 'home_player_Y10', 'home_player_Y11', 'away_player_Y1', 'away_player_Y2', 'away_player_Y3', 'away_player_Y4', 'away_player_Y5', 'away_player_Y6', 'away_player_Y7', 'away_player_Y8', 'away_player_Y9', 'away_player_Y10', 'away_player_Y11', 'B365H','B365D', 'B365A' e 'Output' como feature final da classe.

Aplicando esses dados diretamente numa regressão logística, foi atingido uma acurácia de 46.02%, usando k-fold cross validation com k=10.

Como podemos observar, existem duas features de coordenada para cada jogador num time. Podemos transformar estas 22 features em 2, somente transformando-as na feature formation, que contém número de jogadres na defesa, meio de campo e ataque.

Após essa transformação, denominamos esse conjunto de features e dado como dados-cru. As features do dados-cru são: 'country_id', 'league_id', 'season', 'stage', 'date', 'match_api_id', 'home_team_api_id', 'away_team_api_id','formation_home', 'formation_away', 'B365H', 'B365D', 'B365A' com 'Output' como feature final da classe.

Utilizando regressão logistica, os dados crus apresentaram acurácia média de 46.53%, using a k-fold validation with k=10.

5.1. Feature Engineering e Preração dos dados

Como parte do feature engineering os jogos das 5 primeiras rodadas não foram utilizados para como exemplos, jogos a serem descritos pelas features propostas. O motivador para tal descarte foi o fato de acreditar que nesses jogos os times estão em processo de adaptação, tanto em relação ao campeonato, quanto aos seus novos companheiros, fato que pode tornar os resultados mais aptos a aleatoriedade dos jogos de futebol. Ao desconsiderar este período de adaptação é esperado que a sequência dos resultados seja mais estável em relação ao padrão definido dos times de uma determinada partida. Vale a pena ressaltar que a escolha da quantidade de rodadas a serem descartadas foi completamente empírica.

A cada jogo de uma rodada é levado em consideração todos os jogos das rodadas anteriores daquela equipe. Por

exemplo, em um jogo da rodada 30, os dados presentes na entrada da rede para este jogo, sofrem influências da rodada 1 à rodada 29. Tomando como exemplo o jogo Liverpool vs Manchester United disputado no dia 22/03/2015, pertencente a trigésima rodada da temporada 2014-2015, a Figura 6 demonstra o desempenhos dos times até a vigésima nona rodada.

	J	V	E	D	GP	GC
Manchester United	29	16	8	5	50	26
Liverpool	29	16	6	7	43	30

Figura 6. Desempenho de Liverpool e Manchester United em vinte nove jogos

A Figura 7 e a Figura 8 demonstram o desempenho de Liverpool e Manchester United, respectivamente, ao longo de vinte nove jogos e o desempenho em casa e fora. Na Figura 7 J representa o número de jogos, V o número de vitórias, E o número de empates, D o número de derrotas, GP o número de gols feitos e GC o número de gols sofridos. De forma análoga JC significa jogos em casa, VC vitória em casa etc. O mesmo é válido para F, onde JF representa jogos fora, VF vitórias fora etc.

	M	V	D	L	GP	GC	MH	VH	DH	LH	GPH	GCH
Liverpool	29	16	6	7	43	30	15	8	5	2	24	14

Figura 7. Desempenho detalhado do Liverpool em vinte nove jogos

	M	V	D	L	GP	GC	MA	VA	DA	LA	GPA	GCA
Manchester United	29	16	8	5	50	26	14	4	7	3	17	16

Figura 8. Desempenho detalhado do Manchester United em vinte nove jogos

Para gera estes índices é necessário fazer consultas na tabela Matches e calcular estas taxas.

Também extraí dados das tabelas Team_attributes, Team, Player e Player_Attribute.

Com isso contruí features que eu optei por separar em classes: Base, gols, estilo de jogo e fifa_ratings.As features que compõem cada classe são demonstradas na figura 9.

Ao todo somam-se 97 features propostas a partir do feature engineering feito com 8791 observações para este conjunto completo.

Na proxima seção essa features serão avaliadas quando aplicadas a regressão logística, k-neighbors, Random Forest Classifier e Gradient Boosting Classifier.

```
play_style = ['h_buildUpPlaySpeed',
       'h_buildUpPlaySpeedClass', 'h_buildUpPlayDribblingClass',
       'h_buildUpPlayPassing', 'h_buildUpPlayPassingClass',
       'h_buildUpPlayPositioningClass', 'h_chanceCreationPassing',
       'h_chanceCreationPassingClass', 'h_chanceCreationCrossing',
       'h_chanceCreationCrossingClass', 'h_chanceCreationShooting',
       'h_chanceCreationShootingClass', 'h_chanceCreationPositioningClass',
       'h_defencePressure', 'h_defencePressureClass',
       'h_defenceAggression', 'h_defenceAggressionClass',
       'h_defenceTeamWidth', 'h_defenceTeamWidthClass',
       'h_defenceDefenderLineClass', 'a_buildUpPlaySpeed',
       'a_buildUpPlaySpeedClass', 'a_buildUpPlayDribblingClass',
       'a_buildUpPlayPassing', 'a_buildUpPlayPassingClass',
       'a_buildUpPlayPositioningClass', 'a_chanceCreationPassing',
       'a_chanceCreationPassingClass', 'a_chanceCreationCrossing',
       'a_chanceCreationCrossingClass', 'a_chanceCreationShooting',
       'a_chanceCreationShootingClass', 'a_chanceCreationPositioningClass',
       'a_defencePressure', 'a_defencePressureClass',
       'a_defenceAggression', 'a_defenceAggressionClass',
       'a_defenceTeamWidth', 'a_defenceTeamWidthClass',
       'a_defenceDefenderLineClass']

fifa_ratings=['h_avg_height', 'h_avg_weight',
       'a_avg_height', 'a_avg_weight', 'h_overall', 'h_potential', 'h_def',
       'h_mid', 'h_att', 'a_overall', 'a_potential', 'a_def', 'a_mid',
       'a_att']
base= ['season', 'stage','league_id','a_date', 'h_date','B365H',
       'B365D', 'B365A', 'formation_h', 'formation_a']

goals = ['home_V', 'home_D', 'home_E', 'home_GF', 'home_AVG_GF', 'home_GS',
       'home_AVG_GS', 'home_VG', 'home_DG', 'home_EG', 'home_GFG',
       'home_AVG_GFG', 'home_GSG', 'home_AVG_GSG', 'away_V', 'away_D',
       'away_E', 'away_GF', 'away_AVG_GF', 'away_GS', 'away_AVG_GS',
       'away_VG', 'away_DG', 'away_EG', 'away_GFG', 'away_AVG_GFG',
       'away_GSG']
```

Figura 9. Classes com suas respectivas features.

6. Experimentos

Para a avaliação dos algoritmos será utlizada a métrica multi-class log-loss. Esta métrica é defina pela equação 1.

$$logloss = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij} log(p_{ij}) \quad (1)$$

onde N é o número de observações, M é o número total de classes alvo no problema, y_{ij} é uma variável binária que é ativa quando sua predição é compatível com a classe j e p_{ij} é a probabilidade dada pelo seu classificador daquela observação i pertencer a classe j.

Na figura 10 podemos ver o comportamento da multi-class logloss em relação a probabilidade de uma classe.

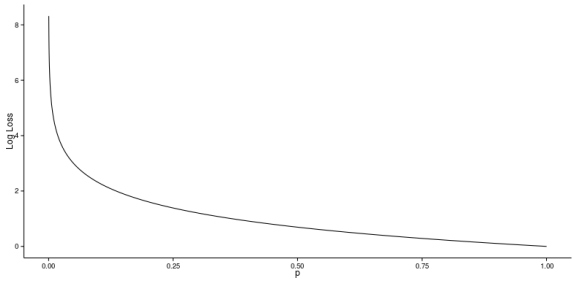

Figura 10. Valor do multiclass logloss vs probabilidade dada a uma classe

Acredita-se que essa métrica seja melhor para o problema, visto que ele leva mais em consideração a probabilidade dada para a classe do que o acerto em si. Como futebol é

um esporte muito suscetível a aleatóriedades, acredito que essa métrica seja mais adequada do que a acurácia.

Todos os experimentos apresentados nesta seção foram realizados usando kfold cross validation com k=10.

Foram feitas uma normalização min-max, uma gaussiana que proporciona media 0. Foram criadas 3 regressões logisticas, uma para cada grupo de dados para testar se a normalização dos dados afetaria o desempenho. Os dados com normalização min-max apresentaram um logloss de 0.9874, normalização gaussiana 0.9916 e os dados sem normalização 0.9814. Logo descartamos as normalizações no dado.

Agora vamos decidir qual o valor de k, no algoritmo k-neighbors. O gráfico da figura 13 demonstra que para k=100 atingimos já o valor minimo para o logloss. No gráfico é dado o logloss invertido.

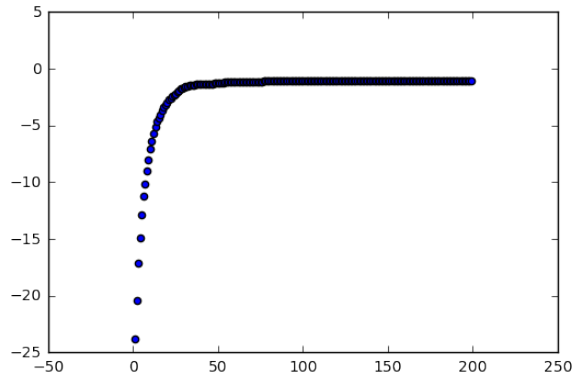

Figura 11. Logloss em função do número de vizinhos

Para k = 100, o modelo apresentou um logloss de 1.090.

Para o algoritmo de random forest, aplicamos uma k-fold cross validation para descobrir o melhor valor possível para o número de árvores. O valor de profundidade máximo das árvores foi fixado em 3. O gráfico da figura **??** demonstra que para k=100 atingimos já o valor minimo para o logloss. No gráfico é dado o logloss invertido.

Podemos observar que temos o menor valor de logloss para k=1000, que foi igual a 0.9850.

O algoritmo Gradient Boosting não teve uma performance satisfatória. Ao fazer o k fold para determinar o melhor valor de estimadores, deu para verificar seu péssimo rendimento. A figura **??** demonstra o rendimento. O motivo pelo pouco alcance é que este algoritmo demora muito para ser executado. Como o desempenho foi péssimo comparado aos outros, já terminei a execução.

Este comportamento é até compreensível, visto que como ele cria uma penalidade pelos erros cometidos pelo último estimador, isso pode atrapalhar muito a predição para o problema do futebol, visto que é um esporte muito suscetível a aleatoriedades.

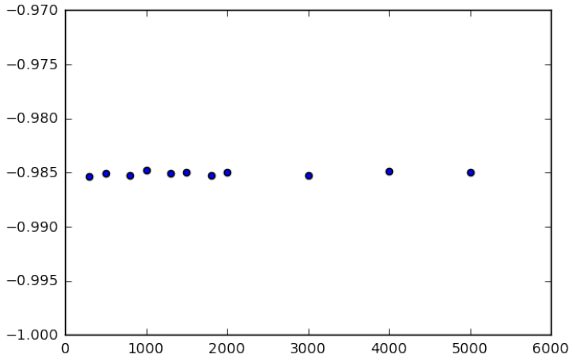

Figura 12. Logloss em função do número de árvores

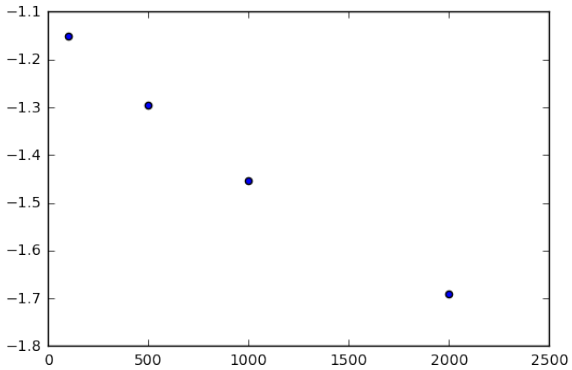

Figura 13. Logloss em função do número de estimadores para o algoritmo Gradient Boosting.

A tabela 1 apresenta o logloss médio dos modelos apresentado acima.

Tabela 1. logloss médio dos algoritmos usando os dados com feature engineering

	logloss médio
Regressão Logística	0.9814
K-Neighbors	1.090
Random Forest	0.9850
Gradient Boosting	1.170

Vamos agora testar os 3 modelos com melhores médias com as diferentes combinações de classes de features. A tabela 2 apresenta o logloss médio dos algoritmos com as combinação das classes de features.

Como podemos observar na tabela 2 a classe Play_Style aumenta o logloss médio dos algoritmos. Podemos observar também que o algoritmo Regressão Logística apresentou um melhor desempenho médio.

Tabela 2. logloss médio de cada algoritmo para cada classe de entrada;

	Reg. Log.	K-NN	R. Forest
Base	0.9735	1.0364	1.0278
Goals	1.0331	1.0378	1.063
Ratings	0.9849	0.9940	1.0097
Style	1.0061	1.0644	1.0883
Base + Goals	0.9741	1.0494	0.9939
Base + Ratings	0.9733	1.0090	0.9931
Base + Style	0.9793	1.0700	1.0034
Goals + Ratings	0.9868	0.9926	0.9974
Goals + Style	1.0390	1.0400	1.0446
Ratings + Style	0.9981	1.0239	1.0036
Base+Goals+Rating	0.9747	1.0263	0.9902
Base+Goals+Style	0.9816	1.0604	0.9930
Base +Rating + Style	0.9802	1.0478	0.9921
Goals + Rating + Style	1.0007	1.0186	1.0006

Partindo para o conceito da filtragem dos dados apresentado por Cheng et al. (Cheng et al., 2003), separei os dados em mandante forte quando a diferença de overall médio do mandante é superior ou igual a 3 unidades em relação ao visitante. Caso o overall do visitante seja maior ou igual ao mandante em 3 unidades esses jogos são filtrados como visitante forte. Se não entrar em nenhum desses filtros, entra para o filtro de jogos com equipes em niveis iguais.

O algoritmo que apresentou melhor desempenho para o mandante forte foi Regressão Logistica com as features da classe Base. O logloss médio foi de 0.8150.

Para os visitantes fortes o modelo que apresentou melhor desempenho foi Regressão Logistica com as features da classe Base. O logloss médio foi de 1.0036.

Para os jogos em níveis iguais o algoritmo que apresentou melhor desempenho médio foi Regressão Logistica com as features da classe Base. O logloss médio foi de 1.0550.

Separando 33% das observações para o teste e criando o modelo híbrido proposto por Cheng et al. (Cheng et al., 2003), obtemos o 1.2853 como logloss médio.

7. Conclusão

Para predição de jogos de futebol, o algoritmo de regressão linear apresentou melhor desempenho médio.

Como predição de jogos de futebol é uma tarefa muito complexa, consegui consolidar conceitos complexos apresentados em sala de aula.

Como trabalhos futuros, criar sistemas mais simples envolvendo futebol que possam sirvir como features para a predição de futebol é o próximo passo que irei seguir.

References

Anderson, Chris and Sally, David. *The numbers game: Why everything you know about football is wrong*. Penguin UK, 2013.

Aslan, Burak Galip and Inceoglu, Mustafa Murat. A comparative study on neural network based soccer result prediction. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*, pp. 545–550. IEEE, 2007.

Cheng, Taoya, Cui, Deguang, Fan, Zhimin, Zhou, Jie, and Lu, Siwei. A new model to forecast the results of matches based on hybrid neural networks in the soccer rating system. In *Computational Intelligence and Multimedia Applications, 2003. ICCIMA 2003. Proceedings. Fifth International Conference on*, pp. 308–313. IEEE, 2003.

de Football Association, Federation Internationale et al. Fifa big count 2006: 270 million people active in football. *Retrieved February*, 20:2012, 2007.

Elo, Arpad E. *The rating of chessplayers, past and present*. Arco Pub., 1978.

kaggle. Soccer database, 2016. URL `https://www.kaggle.com/hugomathien/soccer`.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.