UNIVERSIDADE FEDERAL DO ABC

BACHARELADO EM NEUROCIÊNCIA

Atividade de Pesquisa Científica

PEDRO RICARDO BRONZE

Profº Drº João Ricardo Sato

UFABC

São Bernardo do Campo

2019

Relatório Final para atribuição de conceito

Aluno: Pedro Ricardo Bronze
Orientador(a): Profº Drº João Ricardo Sato
Local: UFABC

Contexto e Motivação:

Com o intuito de analisar dados disponibilizados pelo INEP relacionados ao Índice de Desenvolvimento do Ensino Básico o presente trabalho propõem a implementação de um modelo de regressão baseado no algoritmo de aprendizado de máquina Gradient Boosting Machine. Além disso é realizada análise de forma crítica dos indicadores através das ferramentas disponíveis nas bibliotecas de Python para análise de dados.

Resultados:

O aluno implementou as diversas etapas envolvidas no processo de análise dos dados assim como a implementação do modelo de *GBM,* que consegue explicar a variação da nota do IDEB em função de alguns indicadores principais (*R2 Score* >0.70). Ao fim o aluno conduziu uma análise crítica propondo um cenário descrito pelos dados e apresentando possíveis limitações do trabalho.

Índice

Introdução

Contexto, Motivação e Objetivo

A partir de 2015 dados referentes ao Índice de Desenvolvimento do Ensino Básico (IDEB) passaram a ser coletados e disponibilizados pelo MEC de forma mais ampla (INEP, 2018), sendo expandida para o ensino médio e escolas privadas. Um dos problemas entretanto é a condição que se encontra o ensino básico, por exemplo em 2017 nenhum estado conseguiu atingir a respectiva meta em relação à nota do IDEB. Além disso os índices de abandono e reprovação para o ensino médio logo no primeiro ano são alarmantes, um cenário dramático para a educação do país (INEP - Ministério da Educação, 2018).

É fundamental entender quais as relações entre os indicadores do ensino básico do Brasil e quais relações guardam entre si e entre o IDEB. A geração de *insights* e compreensão das relações que estes guardam entre si é fundamental, para em conjunto com o corpo teórico existente na área da pedagogia agir de forma eficaz propondo alternativas para o atual cenário enfrentado pelo ensino básico do Brasil.

Por meio da criação de um modelo baseado no algoritmo de *Aprendizado de Máquina* Gradient Boosting Machine (GBM) para regressão o presente trabalho visa pesquisar os principais indicadores que poderiam ser utilizados na predição das notas do IDEB. O enfoque sendo o ensino médio, com breve comparação com o ensino fundamental. O aluno busca aproveitou a oportunidade para se familiarizar com a aquisição, limpeza e organização dos dados, assim como as etapas de estatística descritiva inicial, implementação do modelo de GBM, com ajuste de hiperparâmetros e eliminação de características. Ao fim foram discutidos os resultados com breve análise dos principais indicadores envolvidos no modelo e apontadas algumas limitações no trabalho. Então são discutidos os resultados do trabalho comentando principais indicadores envolvidos na análise e a relação destes com a nota do IDEB de acordo com o modelo criado e as possíveis limitações do trabalho.

O presente trabalho foi realizado com extensa utilização de Python (Python Software Foundation. version 3.5) e algumas de suas bibliotecas conhecidas para análise e manipulação de dados como Numpy (der Walt *et al.*, 2011*), PANDAS* (McKinney W., 2010) e *Scikit-Learn* (Pedregosa *et al.*, 2011).

Dados do IDEB

Os dados referentes ao ano de 2017 foram acessados no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP (Inep.gov.br., 2018) que referencia os dados extraídos do Censo Escolar.Os dados dos indicadores do ano de 2017 que estavam disponíveis utilizados foram:

TDI - Taxa de Distorção Idade-Série, HAD - Número Médio de Horas-Aula Diária, IED - Índice de Esforço Docente, IRD - Índice de Regularidade Docente, ICG - Índice da Complexidade da Gestão, DSU - Taxa de Docentes com Ensino Superior, ATU - Número Médio de Alunos por Turma, AFD - Adequação da Formação Docente. A variável de desfecho consiste na Nota do IDEB, disponível no mesmo formato de tabela que os índices restantes.

Inicialmente os dados se encontravam em planilhas separadas, cada escola sendo representada em uma linha enquanto as características estavam representadas em colunas. Enquanto algumas tabelas possuíam mais de 160.000 observações de escolas outras possuíam um número pouco maior que 100.000.

Métrica

A métrica adotada para comparação entre os modelos criados foi a do *Coeficiente de Determinação (R^2 Score)*. Definida da seguinte forma:

$$R^2 = 1 - (SSE/SST)$$

SSE é a *Soma dos Quadrados do Erro Residual*, que basicamente representa a somatória do quadrado da diferença entre a média e os cada um dos valores observados/atuais. Ao passo que SST é a *Soma dos Quadrados Totais,* que equivale a soma do quadrado da diferença entre os valores observados e os valores preditos. O valor geralmente fica compreendido entre 0 e 1, sendo que quanto mais próximo de 1, melhor a regressão consegue predizer os dados. Esse coeficiente captura quanto da variância das variáveis preditoras consegue explicar a variância da variável predita, ou seja, quanto um modelo é bem ajustado à amostra.

Gradient Boosting Machine

Gradient Boosting Machine é um entre os métodos de Assembléias(Ensembles), e como tal se baseia na hipótese de que *weak learners* de tamanho fixo quando juntos podem criar um bom classificador. Este algoritmo pode ser utilizado tanto para problemas de regressão quanto de classificação.

De acordo com Friedman(1999) o algoritmo inicia-se gerando uma função parametrizada simples (*base learner*) treinada nos dados do conjunto de treino, a cada nova iteração é adicionado um novo *base learner* que é treinado nos pseudo resíduos do modelo anterior. Desta forma, utilizando o método de *Gradiente Descendente* o algoritmo itera minimizando a *loss function* ao adicionar novos *weak learners*. Estes *weak learners*, no caso presente se tratam de árvores de regressão.

A implementação do *GBM* na biblioteca do *Scikit-Learn* permite o ajuste de diversos hiperparâmetros que determinam como cada árvore é construída(quantidade, profundidade máxima e número máximo de características por árvore etc) assim como o comportamento do modelo como um todo (*learning rate* e métrica da *loss function*).

Um modelo derivado é o *Stochastic GBM*, no qual subconjuntos das observações são escolhidas (sem reposição) a cada iteração. Adicionando-se aleatoriedade e aumentando a variância entre os *base learners*, porém isto leva a uma menor correlação entre estimadores individuais o que reduz a variância do modelo final. A acurácia tende a aumentar e o tempo de execução do algoritmo tende a diminuir (Friedman, 1999) em relação ao *GBM*. A sua implementação no *Scikit-learn* se dá através da mesma do *GBM*, com a alteração do parâmetro *subsample* que determina o tamanho do subconjunto do conjunto de dados utilizado a cada iteração.

Importância de Características

A biblioteca do *Scikit-Learn* nos permite acessar a importância das características a partir do modelo criado. Em modelos de árvores de decisão a equação para cálculo da importância do nó j é:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

sendo w_j o número de frações de amostras do nó, C_j é a impureza do nó, os subscritos *left(i)* e *right(i)* se referem aos nós-filhos esquerdo e direito, respectivamente. Já a importância da característica i é calculada da seguinte forma:

$$fi_i = \frac{\sum_{j:\text{node j splits on feature i}} ni_j}{\sum_{j \in \text{all nodes}} ni_j}$$

A importância de cada característica é dada pela multiplicação da fração de amostras que a característica contribui para separar pela redução da impureza do nó envolvido.

No caso de modelos de Assembléia como *Gradient Boosting Machine* e *Random Forest* a média da importância de cada uma das variáveis é calculada considerando todas as árvores do modelo (Pedregosa F., 2011).

Feature Elimination/Selection

A etapa de eliminação de características consiste num ponto chave considerando os possíveis diferentes métodos existentes e sua influência sobre as características do modelo em si. Viéses podem ser incluídos sutil e implicitamente nesta etapa comprometendo sobremaneira o modelo, razão pela qual essa etapa deve ser abordada de forma crítica.

Guyon & Elisseeff(2003) apontam algumas questões concernentes ao tema de seleção de variáveis e características. Características redundantes podem conter informação relevante para modelos auxiliando a separar grupos em tarefas de classificação. Similarmente, características sozinhas que aparentam pouca importância, em conjunto com outras podem ser mais informativas do que sua soma separadamente, apontando potencial interação entre variáveis.

Pré-processamento que é executada apenas baseado nos dados pode aumentar o risco de vieses quando comparado com pré-processamento baseado em expertise da área em questão e validação externa.

Nesta etapa tem como principais objetivos:

- Melhorar a acurácia do modelo reduzindo ruído advindo de características com com baixo SNR(Signal-to-Noise Ratio).
- Reduzir custo computacional com um modelo simplificado.
- Revelar possíveis *Hidden Structures* que definam quais dados determinam o comportamento da variável de desfecho.
- Facilitar a compreensão dos dados ao reduzir sua complexidade em número de características.
- Reduzir *overfitting*, o que implica em melhores predições com dados ainda não 'vistos' pelo modelo.

Em resumo, um modelo mais simples é menos custoso e mais fácil de ser entendido, além de definir melhor o conjunto de fatores que regem o comportamento dos dados. *Actionable Data* é o jargão/termo utilizado para referir-se a produção de modelos e dados através dos quais podem-se tomar decisões.

Os 3 métodos mais conhecidos são os seguintes:

Filters

Neste método cria-se um filtro para selecionar variáveis baseado nos critérios adotados pelo usuário. Grau de correlação e informação mútua entre preditoras e variável desfecho, assim como limite de variância para cada variável preditora são alguns dos critérios que podem ser utilizados.

Não tão custosa a ser implementada essa forma de seleção pode ser relevante quando acompanhado de conhecimento do domínio para se eliminar inicialmente características pouco informativas e que adicionam muito ruído a determinados modelos. Ao mesmo tempo, há risco de se gerar um modelo que não considera interação entre variáveis preditoras ao se eliminar baseado em critérios individuais.

O método de Filtro não lida bem com multicolinearidade, que deve ser analisada em etapas prévias de pré-processamento dos dados. Não necessariamente afetando a predição da variável de desfecho, mas possivelmente apresentando resultados inconsistentes sobre as características e o grau de redundância entre si.

Wrappers

No caso adotado no presente trabalho, utilizou-se o algoritmo de *Recursive Feature Elimination.* Neste método, dado um estimador que avalia a importância de cada uma das características, o algoritmo itera eliminando recursivamente as características de menor importância até que o número mínimo desejado de características seja atingido. Desta forma pode-se capturar em mais detalhes as interações entre características gerando assim modelos mais complexos. Outras técnicas consistem em: *Forward Selection e Backward Elimination.*

Embedded

Neste método, a construção do modelo se dá em conjunto com a seleção de características, já avaliando como cada característica auxilia no ganho de informação para um modelo específico. Exemplos de técnicas são as regressões RIDGE e LASSO.

Pré-processamento e modelagem

Junção das tabelas

As bibliotecas utilizadas no processo de *Data Wrangling, Exploratory Data Analysis e Data Cleaning* foram essencialmente *NumPy* e *Pandas* para Python (*versão 3.5*).

- Para a junção as características *Região, UF, Nome do Município, Nome da Escola, Código da Escola, Localidade e Dependência Administrativa* foram utilizadas como referências para unir os dados de cada uma das escolas.

- Após o processo de junção das tabelas alguns *sanity checks* foram realizados checando se os dados continuavam consistentes em alguns pontos. Como por exemplo checar valores únicos para *UF* (27, 26 Estados e o Distrito Federal), *Dependências Administrativas* (4) e *Localidades*(2). Assim como checagem de quantidade de missing values por característica e número total de escolas.

- Valores faltantes estavam no formato '-' e foram substituídos por *numpy.nan,* método que permite que um valor ainda que faltante seja considerado numérico, possibilitando o tratamento numérico dos dados da característica em questão.

- Descartou-se 'Código do Município' pois já haveriam variáveis suficientes (*Região, UF, Nome do Município, Nome da Escola, Código da Escola*) para definir e evitar ambiguidade fosse o caso de haverem escolas de mesmo nome.

- Após as etapas iniciais de *Data Wrangling* realizou-se uma breve análise baseando-se em estatística descritiva, onde grande parte das características apresentaram valores acentuados de valores faltantes(>50%).

Uma hipótese levantada para a grande quantidade de valores faltantes foi a de que cada escola contemplaria diferentes etapas de ensino, e sendo assim não haveriam valores a se preencherem para colunas referentes a outras etapas de cada índice.

Os dados referentes aos notas do IDEB por escola estavam segmentados em Anos Iniciais, Anos Finais do Ensino Fundamental e Ensino Médio. Essa segmentação permitiria uma análise mais adequada etapa a etapa para extração de padrões dos dados.

- Os dados foram segmentados da mesma forma para checar a hipótese acima citada. De fato ao unir via *inner-join* a contagem de valores faltantes se reduziu consideravelmente em cada uma das três tabelas. O que pode ser uma evidência de uma das origens dos valores faltantes.

Explorando as novas tabelas geradas, notou-se que para todas as observações nas quais houvessem valores faltantes, a Nota do IDEB estava entre estes valores faltantes. Considerando se tratar da variável de desfecho neste caso não constitui boa prática tentar imputar valores quais fossem.

- Guiado por essa heurística descartaram-se escolas que possuíam tais valores faltantes.

Para os Anos Iniciais do Ensino Fundamental a perda de amostras foi de ~15 mil de um total de ~55 mil. Nos Anos Finais do Ensino Fundamental foi também de ~15 mil, de um total de ~41 mil. O Ensino Médio houve a maior perda percentual, ~10 mil de um total de ~19 mil escolas observadas. A grande quantidade de dados faltantes já constitui um ponto de alarme para possíveis problemas na coleta ou tratamento inicial dos dados pelos órgãos responsáveis.

Em se tratando de granularidades dos índices, estes apresentavam diversas colunas inicialmente referentes a cada uma das séries de cada etapa. O intuito do presente trabalho foi avaliar padrões por etapas, evitando perda de informação demais ao juntar todas as etapas, tampouco se propondo a uma análise em granularidade mais reduzida(série a série).

- Sendo assim, utilizaram-se as colunas referentes às médias de cada índice para cada etapa. No caso do ICG e IRD ambos são disponibilizados por escola como um todo e não por etapa, sem a granularidade dos anteriores. Sendo assim, não foi conduzida nenhum seleção específica quanto à granularidade nestes índices.

- Realizou-se o processo de *One Hot Encoding* com as características *Localidade* (Rural ou Urbana) e *Dependência Administrativa* (Federal, Estadual, Municipal ou Privada). Assim estas variáveis categóricas foram transformadas em numéricas em formato binário.

- A característica *Código da Escola* foi selecionada para utilização como índice de cada uma das observações. As variáveis identificadoras foram ignoradas nas etapas subseqüentes para construção do modelo.

- Foi observado que não existiam Notas do IDEB referentes às escolas privadas (DA_3) duas etapas referentes ao Ensino Médio. Averiguaram-se os dados originais confirmando que esta falta não era fruto da manipulação inicial dos dados.

Ao fim foram organizadas 3 tabelas com 22 características para as etapas do Ensino Fundamental e 23 para o Ensino Médio mais uma coluna referente à Nota do IDEB.

Estatística Descritiva Inicial

Embora descrito sequencialmente, este processo foi realizado parcialmente em paralelo ao citado acima.

Em análise descritiva inicial foi possível averiguar a consistência dos dados em alguns aspectos.

Inicialmente algumas colunas se destacam com valores nulos em 90% das escolas. Estas características tratavam de etapas de ensino como *Educação de Jovens e Adultos(EJA)* e outras categorias como *Turmas Multietapa, Multiseriado ou Correção de Fluxo/Educação Profissional/Educação Especial.* Estas colunas foram descartadas, haja visto o escopo do presente trabalho.

As colunas dos índices que apresentavam a média por etapa apresentaram sempre menores quantidades de valores nulos. O que faz sentido, pois mesmo que faltem dados mais granulares(série a série), as médias serão computadas utilizando o conjunto das colunas restantes para cada índice e para cada etapa.

Uma característica peculiar é que ao analisarmos os índices em sua menor granularidade (série a série) as contagens de observações tendem a diminuir. Essa questão pode ser fruto do índice de desistência que refletiria na quantidade de turmas disponíveis para matrícula, reduzindo a contagem de observações totais.

No caso do IED como se tratam de Níveis (1 a 6) agrupados por etapa de ensino, eles possuem as mesma contagens dentro das respectivas etapas, isso faz sentido, uma vez que quando há uma amostra coletada de uma etapa ela é coletada para todos os níveis. Quanto maior o nível deste índice, maior o esforço do docente, que é calculado indiretamente baseando-se em número de escolas, turnos, alunos e etapas nas quais o docente atua.

Os índices TDI, IED, DSU e AFD são dados em porcentagens o que condiz com seus valores mínimos e máximos, 0 e 100 respectivamente.

Para ATU, temos um mínimo de 1 aluno por turma, e máximos de valores superiores a 200 alunos por turma para alguns casos no Ensino Médio. É possível imaginar turmas deste tamanho, porém podem constituir indícios de coleta inadequada dos dados. Apesar disso, o Intervalo Interquartil (IIQ) se encontra entre 15 alunos e valores que não ultrapassam 40, com médias e medianas muito próximas podemos interpretar que existem poucos outliers que afetem a distribuição destes dados.

No caso do HAD o IIQ revela que para turmas do Ensino Fundamental temos um grande número delas entre 4 e 5 horas-aula diária(had), para o Ensino Médio o valor máximo sobe para pouco mais de 5 had. Embora hajam valores de menos de 1 had, os valores máximos não passam de 16.5 had ainda estão dentro de valores reais para o intervalo de 24h correspondente a um dia.

O IRD com média de 3.03, mediana de 3.05 e 3º Quartil de 3.52 tendo como máximo 5, parece apontar que grande parte das escolas tem uma média consideravelmente baixa de regularidade docente, algo alarmante. Uma das limitações é que o índice não aponta a distribuição por etapa, apenas uma média por escola.

O DSU apresenta médias e medianas crescentes de acordo com as etapas, porém com médias bem abaixo dos valores das medianas, enquanto o 3º Quartil se encontra em 100 (valor máximo) para todas as etapas. Uma possível análise é de que há um conjunto não muito grande de escolas em relação ao total, porém com valores consideravelmente baixos, afetando a média. Apesar disso, o 3º Quartil se encontra em 100% (valor máximo), sendo assim ao menos 25% de todas as escolas possuem 100% dos seus docentes com ensino superior.

O índice AFD é subdividido em 5 Níveis, sendo que o Nível 1 representa a porcentagem de docentes com melhor nível de adequação o Nível 5 a porcentagem de docentes com a formação comparativamente é menos adequada à disciplina que lecionam.

Dos histogramas com distribuições de quantidade de escolas em cada porcentagem para cada nível é possível extrair que no Ensino Médio existem poucas escolas com 0% de docentes com Nível 1 de AFD. O histograma do Nível 3 revela que grande parte das escolas tem baixas porcentagens de docentes neste nível. Além disso existe grande quantidade de escolas com 0% de docentes com Níveis pouco adequados (4,5). Em contrapartida a quantidade de escolas com quase 100% dos docentes no Nível 5 para Anos Iniciais e Finais se destaca. Divergindo consideravelmente do Ensino Médio, os valores de médias, medianas e principalmente 3º Quartis corrobora essa análise.

Primeira Etapa de *Eliminação de Características*

O trabalho de selecionar características iniciou-se em conjunto com a etapa de *Data Wrangling*. As características disponíveis em diversos casos possuíam uma granularidade de séries, sendo que em algumas delas apresentavam colunas referentes a Ensino para Jovens e Adultos(EJA), Multiseriados etc. Mantiveram-se apenas as colunas referentes a totais/médias de cada índice para respectiva etapa, procurando assim por padrões na média e não série-a-série. Sendo assim reduziu-se de um total de aproximadamente 100 colunas para 23 colunas de características, considerando a binarização de *Localidade* e *Dependência Administrativa*.

Implementação do modelo de GBM

Utilizando a biblioteca do *Scikit-Learn* foi implementado o algoritmo de *GBM* com os seguintes hiperparâmetros iniciais:

GradientBoostingRegressor(
n_estimators=100, learning_rate=0.1, max_depth=5,
max_features='sqrt', subsample=1.0,random_state=10,
loss='ls')

Com 100 árvore iniciais, um máximo de características correspondentes a raiz quadrada da quantidade total para a construção de cada árvore. Usando como função a otimizar *regressão dos mínimos quadrados*. Os restantes hiperparâmetros são os padrões da implementação no *Scikit-Learn (versão 0.20.2)*.

Nesta configuração o modelo apresenta um *R2 score* inicial de **0.7010** com o ranking de importância de características apresentado na Figura 1.

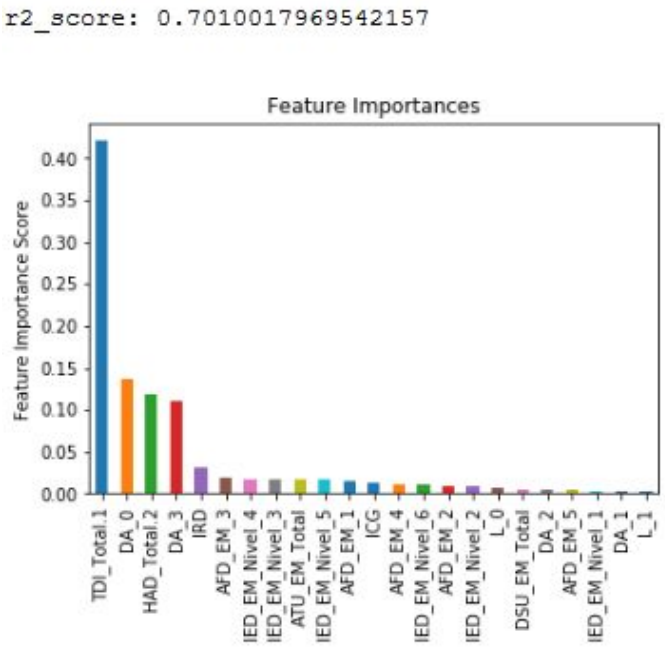

Figura 1 - Ranking de Características para modelo inicial

Ajuste de Hiperparâmetros

Para o ajuste de parâmetros do modelo de GBM o algoritmo de *GridSearch* com validação cruzada foi utilizado. Este método de otimização de hiperparâmetros itera ao longo de um lista de valores passados e busca o modelo com melhor score para esta lista. Para avaliação mais robusta a implementação do *Scikit-Learn .GridSearchCV* (Pedregosa et al., 2011) permite definir um número de *folds* para se realizar validação cruzada (CV) a cada valor da lista do hiperparâmetro escolhido.

A ordem do ajuste foi : *n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, subsample.* Sendo que *max_depth* e *min_samples_split* foram ajustadas em uma mesma grade de iterações. O último ajuste foi feito manualmente reduzindo-se o *learning_rate* enquanto o *n_estimators* foi aumentado proporcionalmente, até que não houvessem mais ganhos significativos. A métrica de *score* para avaliação do ajuste dos hiperparâmetros foi R2. O modelo pós ajuste teve a seguinte configuração de hiperparâmetros:

GradientBoostingRegressor(
n_estimators=1500,max_depth=6, learning_rate=0.005,
subsample=0.8,min_samples_split=5,min_samples_leaf=10,
max_features=3, loss='ls')

Com um subsample de 0.8 obteve-se um *R2 Score* maior, sendo assim segundo a definição de Friedman(1999) se trata de um modelo baseado no algoritmo de *Stochastic Gradient Boosting Machine.*

Avaliação

Após a etapa de ajustes o modelo teve um *R2 Score* de **0.7034**, o gráfico da Figura 2 apresenta a ordem de importância de cada uma das características.

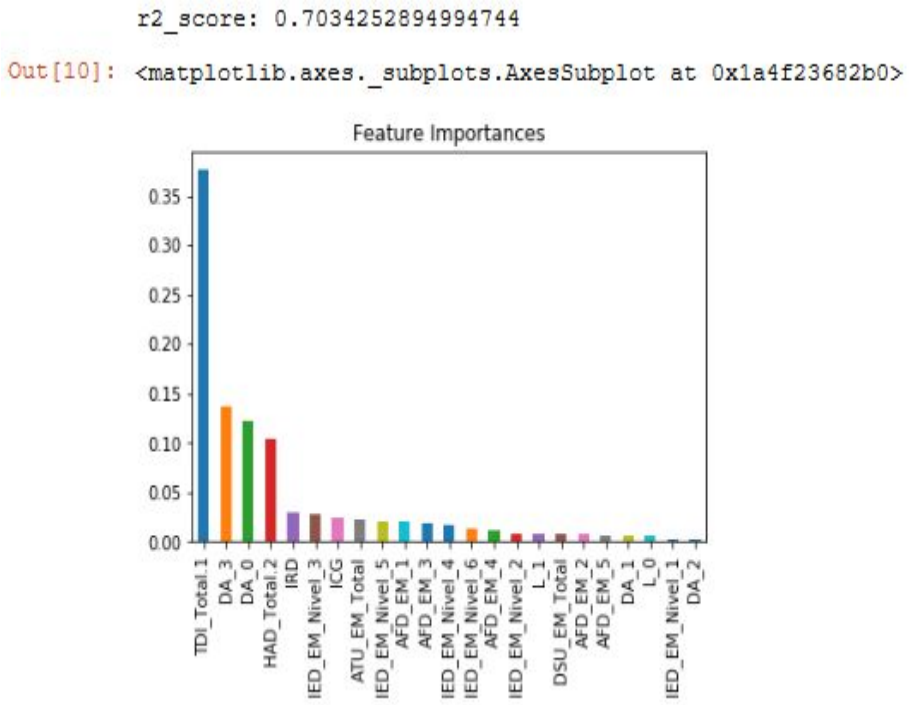

Figura 2 -Ranking de Características após Ajuste de Hiperparâmetros

Segunda Eliminação de Características

A segunda etapa de Seleção de Características se deu após o ajuste do modelo via *GridSearchCV*, desta vez entretanto utilizaram-se os métodos de *Filtro* e *Wrapper*. Para efeitos comparativos utilizando-se a implementação do *Scikit-Learn* para o algoritmo de *RFECV* utilizando os seguintes hiperparâmetros:

RFECV (GBM,min_features_to_select=12,cv=5, step=1)

Definindo assim 12 características foram selecionadas a cada *fold* das 5 etapas de *Validação Cruzada,* sendo que a cada iteração eliminou-se 1 característica. O modelo final apontou que com 19 das 23 características iniciais o *R2 score* foi de **0.7061**, eliminando assim DA_1 , DA_2 , IED_1 e L_1.

Ao utilizar o método de filtro foi requerido que o mesmo número de características fossem selecionadas para efeitos comparativos. A métrica de avaliação para filtragem foi a *mutual_info_regression* implementada pela biblioteca *Scikit-Learn.* Consiste numa medida de dependência de variáveis baseada na quantidade de informação que se tem de uma variável ao se observar uma segunda variável, no caso cada uma das características em relação a variável de desfecho. Esta função utiliza métodos não-paramétricos que se baseiam na estimação de entropia das distâncias dos k vizinhos próximos (Kraskov A. et al., 2004).

Obteve-se um *R2 score* de **0.7044** eliminando por outro lado DSU , DA_2 , IED_1 e AFD_EM_2.

Enquanto pelo método *RFECV* foram selecionadas DSU e AFD_EM_2, o filtro de informação mútua selecionou L_1(Localidade Urbana) e DA_1 (Dependência Administrativa Federal).

A importância das características após a eliminação de características de acordo com cada um dos métodos propostos é apresentada nas Figura 3 e Figura 4.

```
r2_score: 0.7061381793782047
```

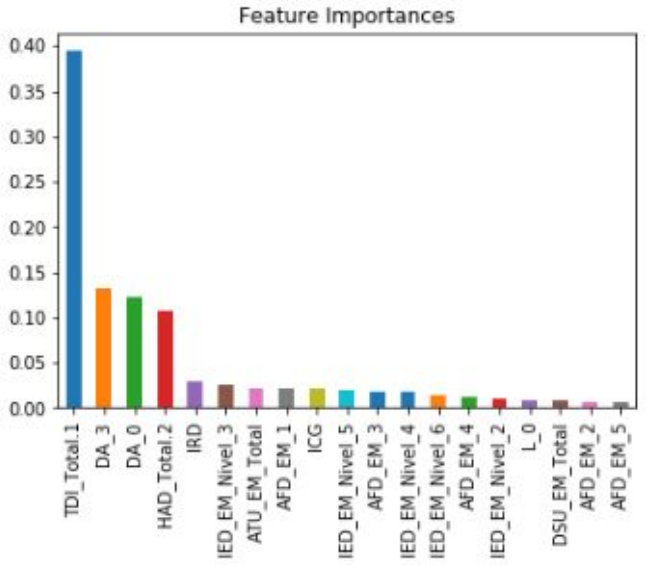

Figura 3 - Ranking de Características após eliminação via RFECV

```
r2_score: 0.7044092791488135
Text(0, 0.5, 'Feature Importance Score')
```

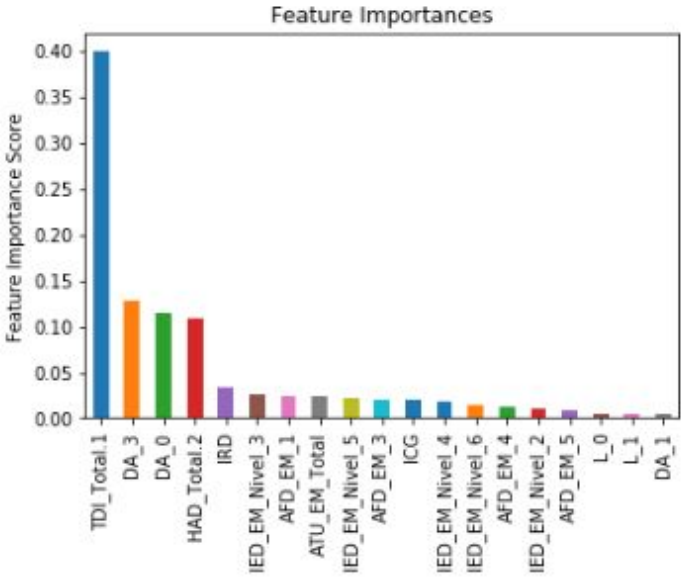

Figura 4 - Ranking de Características após eliminação via Filtro (Informação Mútua)

Discussão

Comparação de Modelos

O modelo inicialmente atribui mais de 0.4 de importância para *TDI*, enquanto as variáveis DA_3(Dependência Administrativa Privada) e IED_3 (Índice do Esforço Docente Nível 3) não parecem ter uma importância que chegue próximo daquela. Ajustando os hiperparâmetros já nota-se uma alteração sutil com DA_3 agora ranqueada em 2ª posição e IED_3 de 8ª para 6ª. Outro detalhe é que TDI tem sua importância reduzida para menos de 0.4 após o ajuste e se mantendo após a fase de RFE. Refletindo que o modelo passa a utilizar o que julga ser informação contida em outras características para melhorar sua predição. AFD_3 passa de 6ª para 11ª posição mostrando que conforme se eliminam características de menor importância, pode haver uma reestruturação da importância das restantes.

RFECV

As características eliminadas pelo método *RFECV* foram DA_1, DA_2, IED_1 e L_1, as 3 primeiras estavam entre as 4 menos importantes no modelo de hiperparâmetros ajustados. Porém L_1 ocupa a 8ª menor posição, não obstante é eliminada ainda considerando que DSU, AFD_2 e AFD_5 a priori apresentavam-se menos importantes. L_1 (Urbana) e L_0 (Rural) se tratam de características binarizadas (*dummy*) assumindo valor 1 na coluna que representa a *Localidade* da escola e 0 na outra coluna. O método parece capturar a informação de que apenas uma das variáveis de *Localidade* seria suficiente ao optar por eliminar L_1. Além disso, L_0 parece assumir a posição de L_1 e continua com uma importância maior do que DSU, AFD_2 e AFD_5.

Um outro método utilizado na eliminação de características é o método de *Variance Threshold* (Limiar de Variância), no qual eliminam-se variáveis com variância abaixo de um determinado limiar. ATU (Número Médio de Alunos por Turma) apresenta uma das menores variâncias entre as todas as características (4ª menor no total das 23 características), ainda assim está em 7ª posição em importância. A aplicação de um filtro deste tipo poderia ter eliminado informações importantes contidas nessa característica.

Filtro

O filtro baseado na informação mútua entre características e notas preditas do IDEB eliminou as características IED_1, DA_2, AFD_2 e DSU. Uma análise importante a se fazer é tanto avaliar as características eliminadas quanto as que não foram eliminadas. O filtro, por exemplo, consegue avaliar que L_1 e L_0 guardam relação com a Nota do IDEB, porém não consegue captar que ambas se referem a uma única característica binária. Sendo assim o filtro opta por manter estas variáveis no lugar de eliminar uma delas e selecionar um conjunto de características mais informativo para o modelo.

Analisando características

A *Taxa de Distorção Idade-Série* (TDI) tem forte (0.82) correlação negativa com as notas do IDEB. Escolas de Ensino Médio com maiores taxas de alunos defasados em relação a série que deveriam estar cursando na sua idade tendem a ter pior desempenho no IDEB. O *número médio de Horas-Aula Diária* (HAD), apresenta média (0.32) correlaç com TDI. Sugerindo uma leve tendência de escolas com menor HAD possuírem maiores valores de TDI. No gráfico de matriz de pontos apresentado na Figura 5, ao observarmos a coluna HAD_total.2 (HAD) em relação a TDI_Total.1 (TDI) pode-se perceber que nas escolas com maior HAD altos valores de TDI não são abundantes.

Na figura 7 o gráfico de distribuição de notas também aponta que não parecem haver escolas com menos de 4 horas-aula diária (HAD) com desempenho que se destaca. Escolas com Notas do IDEB previstas mais altas em grande conta possuem HAD maior do que 4, sendo que o intervalo entre 4 e 6 horas contêm grande parte destas escolas.

Parece existir um aglomerado de escolas que possivelmente se refere ao ensino integral entre 8 e 10 horas. Este agrupamento possui amplitude menor de notas, com mínimos mais altos e máximos mais baixos em relação ao agrupamento compreendido 4 e 6 horas.

Na figura 7, no gráfico dos valores preditos para as Notas do IDEB em relação ao *Índice de Regularidade Docente* (IRD), podemos observar que escolas com valores de IRD até pouco mais de 2 não apresentam boas notas preditas pelo modelo. A distribuição de máximos parece seguir uma linha sendo que as notas mais altas se encontram em escolas com IRD próximo de 4, sendo que o máximo do índice é 5. Apesar de IRD parecer influenciar as notas preditas do IDEB, a correlação destas variáveis é de 0.21 (Figuras 6 e 7).

DA_0 e DA_3 se referem respectivamente às escolas Privadas e Estaduais, e no Pode-se constatar a correlação mútua entre estas duas variáveis e em menor grau TDI (0.35/-0.36), IRD (-0.26/0.26) e em maior grau a nota prevista do IDEB (-0.67/0.68), (Figura 6). Deixando clara a tendência de escolas estaduais apresentarem maior TDI, menor IRD e menor nota do IDEB, ao contrário das privadas.

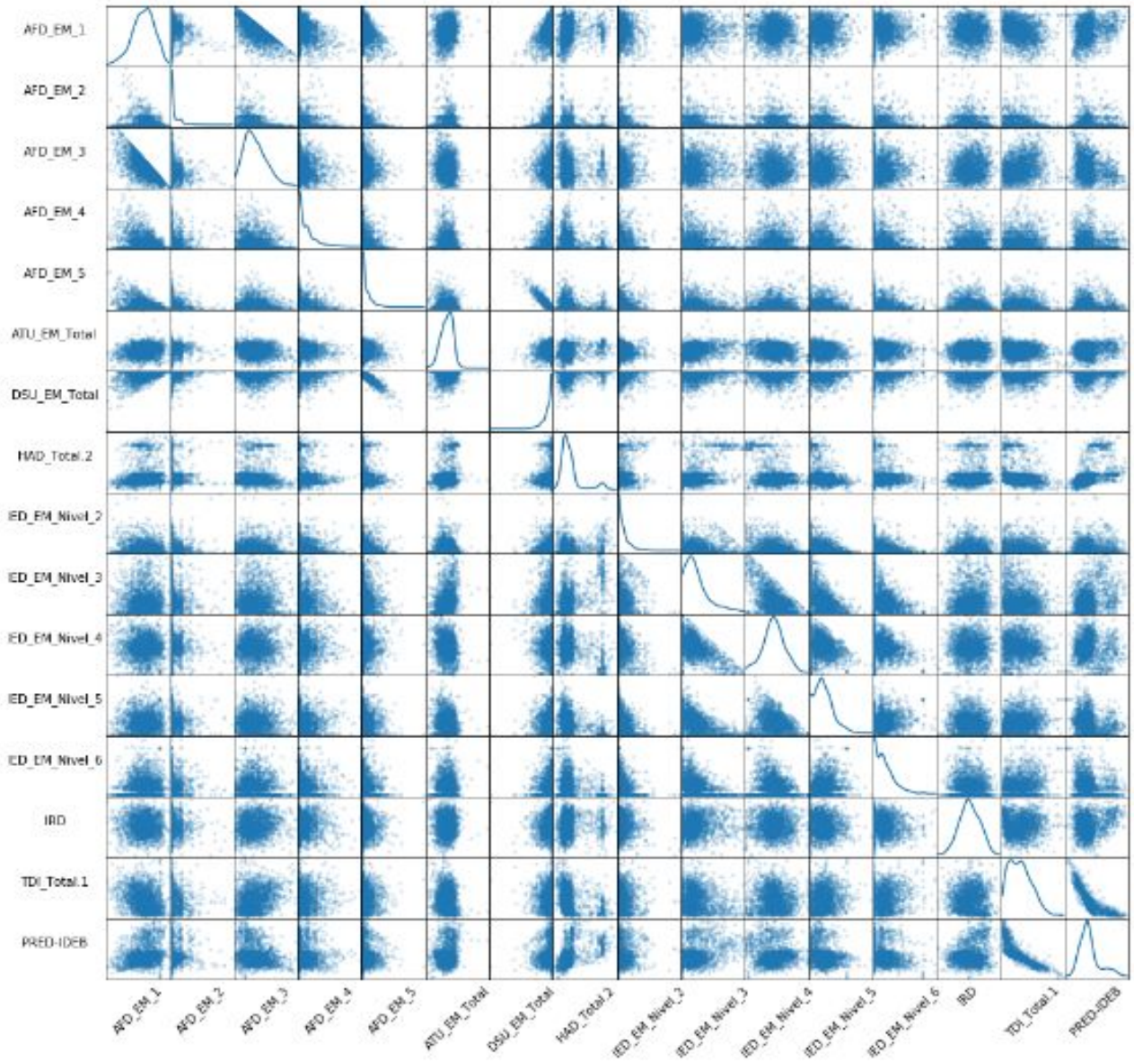

Figura 5 - Scatter Plot Matrix de Variáveis pós-eliminação via RFECV

HAD apresenta uma correlação de 0.47 em relação às notas preditas do IDEB, evidenciado na Figuras 6 e 7.

A variável *Índice de Esforço Docente* (IED) está distribuída entre 5 níveis, sendo assim sua avaliação é um pouco mais delicada do que as outras. O modelo elenca o nível 3 (IED_3) como tendo um nível de importância considerável para o modelo. Além disso, IED reflete uma medida de com quantos alunos, turmas e turnos o docente trabalha, não o quanto ou a natureza do esforço do mesmo.

O modelo embora tenha pouco ganho no desempenho em termos de *R2 Score*, da sua implementação inicial até RFECV, parece atribuir maior importância a ATU e AFD_1. Apontando que estas características parecem ter certo grau de informação, merecendo devida atenção.

Entre escolas com ATU reduzido há maior número de notas baixas preditas do IDEB. Sendo que abundam escolas com maiores notas entre as com ATU entre pouco menos de 20 e pouco mais de 40 alunos por turma em média.

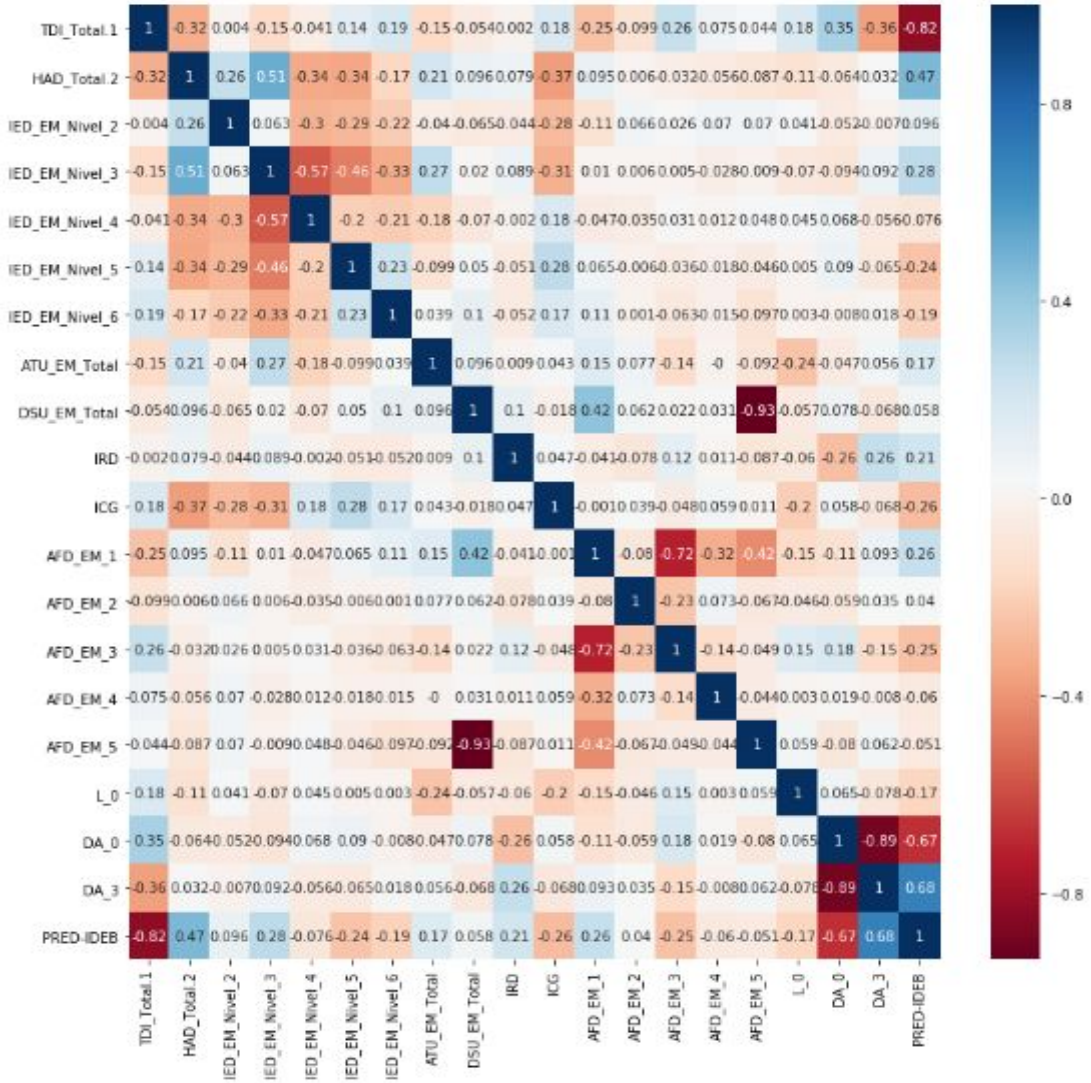

Figura 6 - Matriz de correlação entre características pós RFECV

As distribuições de notas preditas em relação a AFD_1 e a IRD se assemelham. Em ambas os valores máximos acompanham o crescimento destes índices. O Nível 1 representa o número percentual de docentes em uma escola com o melhor nível de adequação. O desempenho das escolas nas quais o percentual de docentes em Nível 1 é inferior a 25% parece ser pior, enquanto que os máximos parecem acompanhar o crescimento percentual deste índice. Os valores máximos se encontram em escolas onde o AFD_1 é próximo de 100%. Para as demais características que tem por base AFD, valores menores de notas são mais corriqueiros com maiores valores destas características.

Embora ambas AFD e IED estejam distribuídas em um conjunto de níveis(colunas), AFD parece medir algo mais concreto, observando apenas a formação docente. O que parece ter uma influência mais perceptível no modelo ao observarmos a Figura 7.

Embora L_0 apresente pouca importância no modelo checando a distribuição de notas preditas em relação a localidade nota-se que escolas urbanas tem distribuição mais ampla atingindo valores mais altos do IDEB em relação às rurais (Figura 7).

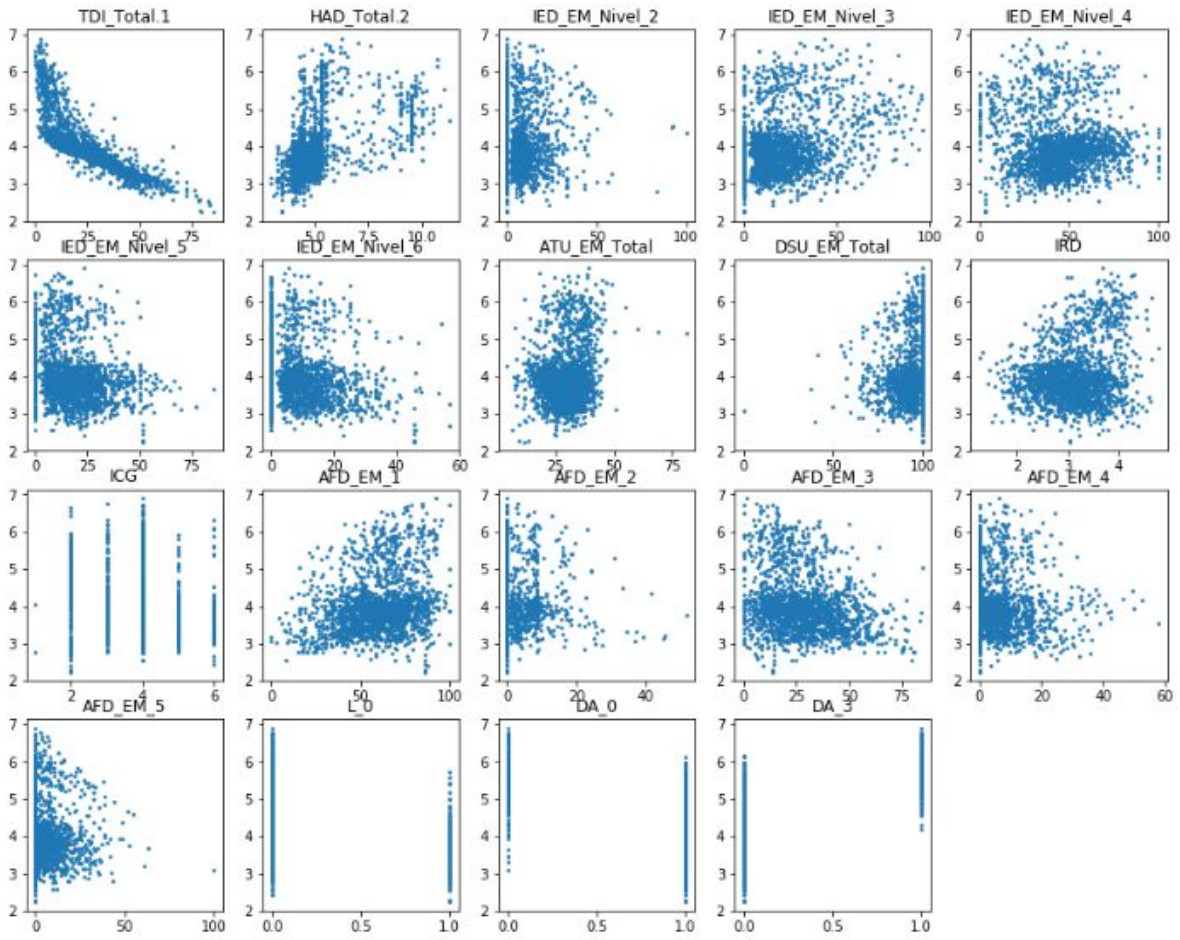

Figura 7 - Notas Preditas do IDEB em relação a cada característica selecionada

Poucas escolas apresentam DSU abaixo de 75% que apresenta correlação insignificante (0.058) em relação a nota predita do IDEB (Figura 6). Não são preditas escolas com notas altas onde o percentual de Docentes com Ensino Superior (DSU) é reduzido. Um valor de DSU mais alto não é garantia de maiores valores do IDEB, porém estes valores mais altos não se apresentam onde o valor de DSU é reduzido.

Ainda que DSU pareça conter informação relevante, existe grande quantidade de escolas com alto índice de DSU (próximo de 100%), sendo que em alguns casos tem valores extremamente baixos e outras consideravelmente altas para as predições da nota do IDEB. As árvores do modelo portanto não conseguem utilizar essa característica para reduzir a impureza dos nós, para o valor de 100% DSU parece que temos muitos valores e muito distribuídos.

Ensino Fundamental

Os scores obtidos na implementação do *GBM* para as etapas do Ensino Fundamental de início foram consideravelmente menores que os scores para o Ensino Médio. Anos Iniciais com um *R2 Score* de *0.5844* e Anos Finais o *R2 Score* é de *0.5274*, revelando que nestes casos o modelo criado não consegue explicar de forma tão clara a variação da nota do IDEB em função das variâncias das características do modelo.

As características de maior importância são o TDI, os extremos da adequação docente (AFD_1 e AFD_5), assim como o DSU, HAD e a característica referente a escola ser Urbana (L_0) (Figura 8). Há discrepâncias fundamentais entre as características que parecem explicar o modelo no Ensino Fundamental em relação ao Ensino Médio que prioriza as dependências administrativas (Federal e Privada), IRD e IED_3.

Quantitativamente parece que no Ensino Fundamental, estes modelos não conseguem fazer uso da informação presente no restante das características além de TDI, que apresenta importância de mais de 55% em ambos os casos. O modelo depende muito das informações referente a distorção idade-série para predizer a nota. em termos de valores iniciais estes modelos já apresentam valores consideravelmente maiores do que o apresentado para no modelo do Ensino Médio. Uma questão que pode afetar o modelo é o fato de que não há registros das notas do IDEB para as escolas privadas desta etapa.

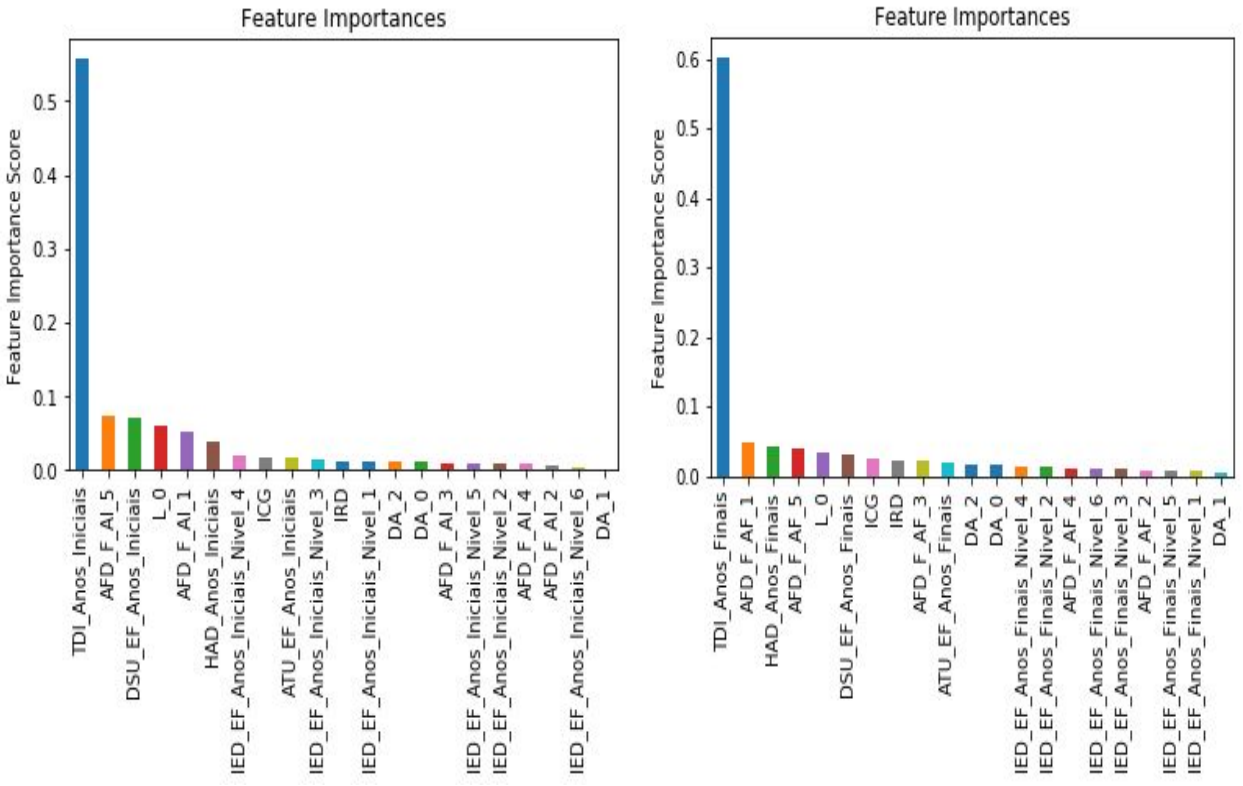

Figura 8 - Ranking de Características para Anos Iniciais (esquerda) e Finais (direita), na primeira implementação do modelo de GBM.

Limitações

Considerando que o presente trabalho visou também familiarizar o aluno com as principais etapas da análise de dados e implementação de um modelo de aprendizado de máquina algumas limitações são claras. Diversas questões surgiram ao longo das etapas da análise, uma avaliação mais criteriosa apresentaria melhores soluções para as adotadas nas épocas que as questões surgiram.

Nas etapas iniciais ao deparar-se com uma quantidade considerável de *missing values* em se tratando da variável de desfecho optou-se por eliminar as observações em questão. Neste caso mais da metade das observações do Ensino Médio foram perdidas, o que pode ter ocasionado perda considerável de informações relevantes para o modelo. Por outro lado, é possível que com menos amostras o modelo tenha conseguido reduzir seu erro de predição, porém apenas para este grupo de escolas, sendo assim com baixa generalização. Ademais a amostra pode ter se tornado mais enviesada caso haja um padrão do perfil de escolas sem dados da nota do IDEB.

Após aprofundar estudos sobre a divisão de subconjuntos de treino, validação e teste se tornou mais clara para o aluno. Inicialmente o modelo contava apenas com conjunto de treino e de teste. Isto pode ter levado ao fenômeno de *data leakage*, quando dados do conjunto de teste são *vistos* pelo modelo antes do teste final, levando assim a estimativas mais otimistas. Esse fenômeno é apontado no livro “data mining (Nisbet R, 2009. Chapter 20) como um dos 10 erros elementares na mineração de dados. Por outro lado, desde o começo o modelo já apresenta um *R2 Score* mais alto, logo mesmo com a ocorrência de *data leakage* o modelo não se torna consideravelmente otimista.

Um ponto não abordado aqui porém extremamente importante se refere a taxa de abandono e como esta poderia estar relacionada com outras variáveis.

Outro ponto importante a se destacar é que em modelos onde é utilizado o ajuste de hiperparâmetros, a melhor prática é a utilização de *Nested Cross Validation*. Não até ser tarde o aluno descobriu isto e a importância da realização do procedimento desta forma para reduzir as chances de sobreajuste do modelo.
A distribuição de notas do IDEB deixa claro que notas altas (>6) não são tão frequentes (Figura 9). Isto pode ser um caso similar ao que seria um problema de classificação de *classes desbalanceadas*. Entender padrões a partir destas poucas observações sem algum método de *oversampling* ou *undersampling*, pode ser problemático e em alguns casos levar a sobreajuste nos casos onde há pouco consenso entre as árvores (*base learners*)(Torgo et al., 2015). Isto poderia inclusive constituir uma das hipóteses para se explicar os casos de alguns resíduos acima de 2 pontos de desvio dos valores reais, para algumas escolas com 5 ou mais pontos (Figura 9).

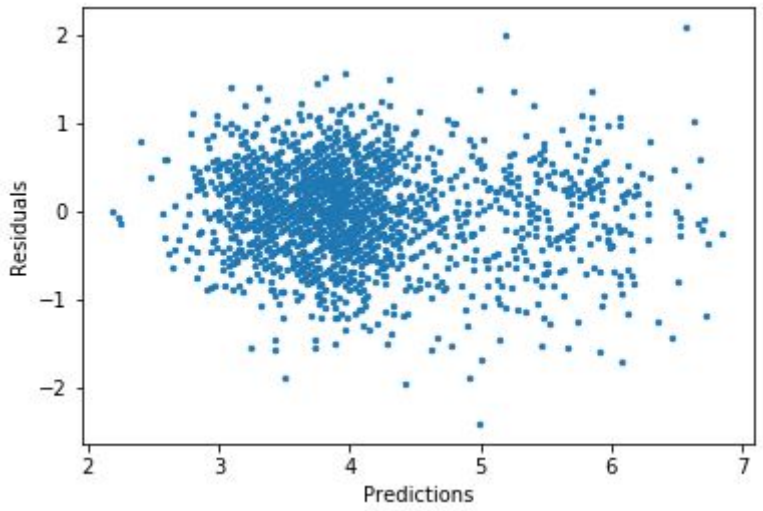

Figura 9 - Valores da nota do IDEB (Preditos e Resíduos)

Por fim as variáveis binarizadas Dependência Administrativa (DA) embora apresentem alta importância para o modelo e tenham seu valor preditivo, não constituem características nas quais poderiam-se gerar dados acionáveis num primeiro momento.

Conclusões

O algoritmo *Gradient Boosting Machine* ao definir uma barreira de decisão não-linear consegue capturar melhor as relações não-lineares entre as características advindas dos dados fornecidos pelo *INEP*. O modelo criado atingiu um *R2 Score* de 0.7061 evidenciando que aproximadamente 70% da variação da nota do IDEB parece conseguir ser explicada pela variação das características apresentadas. Como abordado na seção de ***Limitações***, o fenômeno de *data leakage* pode ter ocorrido dado que a separação em subconjuntos de Treino, Validação e Teste não foi feita da melhor forma.

Apesar disso, o *R2 Score* no primeiro contato com o conjunto de teste já é superior a 0.7000, portanto ainda que tenha ocorrido *data leakage* em etapas posteriores, a estimativa não se torna muito mais otimista que a inicial. Por um lado o *R2 Score* do modelo não parece melhorar muito em face das etapas de *Eliminação de Características*, por outro consegue-se reduzir em quase 20% a quantidade de características necessárias. Auxiliando na compreensão do comportamento dos dados do IDEB enquanto se reduz o gasto com processamento de dados pouco informativos.

Em se tratando das características propostas para criação do modelo, ao passo que nenhuma determina inequivocamente a nota do IDEB, na Figura 7 é evidente que existem tendências e intervalos na distribuição das notas do IDEB em relação a cada característica.

Com uma importância alta da TDI, podemos inferir que a *Taxa de Distorção Idade-Série* foi responsável por quase 40% da redução de impureza média, do conjunto de árvores criadas. A soma entre as Dependências Administrativas, Privada (DA_3) e Estadual (DA_0), o *número médio de horas-aula diária* (HAD) e o *Índice de Regularidade Docente* (IRD) equivale a aproximadamente o valor da TDI. Escolas com altas taxas de TDI coincidentemente tendem a deter notas mais baixas do IDEB.

Lembrando que o modelo não captura relações de causalidade, porém indica possíveis caminhos a se seguir para aprofundar o estudo da interação destas variáveis. Nesse cenário um ponto a se observar é que embora a *Dependência Administrativa* não constitua uma variável acionável, HAD e IRD, assim como ATU, que é a próxima variável mais importante no modelo, representam características possíveis de serem acionadas.

Nesta conjuntura Escolas Estaduais com menos de 4 horas-aula diária em média (HAD), tendem a ter uma nota do IDEB predita menor em função de maiores *Taxas de Distorção Idade-Série* (TDI) e até certo grau exibem sensibilidade a uma menor regularidade docente (IRD). Este grupo de escola parece representar um grupo de risco de níveis consideravelmente baixos de desempenho no IDEB.

Referências

Cawley G.C., Talbot N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation (2010) *Journal of Machine Learning Research*.11(Jul):2079−2107.

der Walt S.,Colbert S.C. and Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation, (2011) *Computing in Science & Engineering*, 13, 22-30.

Friedman J. H. (2002) Stochastic gradient boosting. *Computational Statistics and Data analysis*.38 , 367.

Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection (2003) *Journal of Machine Learning Research* (Mar):1157-1182.

Inep.gov.br. (2018). *Indicadores Educacionais - INEP*. [online] Disponivel em: http://inep.gov.br/indicadores-educacionais. Acessado em: 15/12/2018.

INEP (2018). *IDEB – Planilhas com resultados por escola já estão disponíveis - Artigo - INEP*. [online] Disponível em: http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/id/1511536 . Accessado em: 13/12/2018.

INEP - Ministério da Educação "RESULTADOS DO ÍNDICE DE DESENVOLVIMENTO DAEDUCAÇÃO BÁSICA–IDEB 2017" 2018. Disponível em: http://download.inep.gov.br/educacao_basica/portal_ideb/planilhas_para_download/2017/IDEB2017_APRESENTACAO_final.pdf. Acessado em: 01/03/2019

Kraskov A. , Stogbauer H., Grassberger P., Estimating mutual information. (2004) *Phys. Rev. E* 69, 9.

McKinney W. Data Structures for Statistical Computing in Python (2010) *Proceedings of the 9th Python in Science Conference*, 51-56.

Nisbet R, Elder J, Miner G. Handbook of statistical analysis and data mining applications. 1st ed. Amsterdam: Academic Press/Elsevier; 2009.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.. Scikit-learn: Machine Learning in Python (2011) *Journal of Machine Learning Research*, 12, 2825-2830

Python Software Foundation. Python Language Reference, version 3.5. Disponível em: http://www.python.org. Acessado em: 20/12/2018

Torgo L., Branco P., Ribeiro R.P., Pfahringer B. Resampling strategies for regression. (2015), *Expert Systems*, 32, 465– 476.