TL;DR

Esse desafio consiste em criar um modelo de Machine Learning que preveja o preço de casas baseado em dezenas de características das mesmas.

Os dados utilizados são públicos e fornecidos pela competição *House Prices - Advanced Regression Techniques,* do Kaggle.

O objetivo é chegar o mais próximo possível do score de 0.11230, sendo que qualquer resultado acima de 0.14460 é um resultado ruim. (deixando claro que quanto menor o score, melhor o modelo).

O melhor resultado que obtivemos até o momento foi 0.12365. Essa performance nos possibilita entrar no top 10% da *leaderboard* da competição (considerando apenas os resultados legítimos).

O projeto será atualizado regularmente com o objetivo de melhorar nosso score.

Meus sinceros agradecimentos à Pedro Marcelino, Aurélien Geron e Gabriel Atkin, fontes essenciais para o bom resultado desse projeto.

Descrição e informações sobre o projeto:

Esse projeto foi feito com o objetivo de pôr em prática meus estudos envolvendo as etapas de um projeto de Machine Learning, como limpeza de dados, transformação de atributos e criação e avaliação de modelos. Os dados são fornecidos por uma famosa competição do Kaggle, a *House Prices - Advanced Regression Techniques,* cujo objetivo é criar um algoritmo que preveja o preço de casas na cidade de Ames, Iowa, nos Estados Unidos.

Ainda que esse Dataset seja utilizado exaustivamente por aqueles que estudam Machine Learning, decidi utilizá-lo neste meu primeiro projeto de regressão publicado por alguns motivos: o primeiro deles é que eu já havia feito alguns experimentos com esse Dataset nos primeiros meses em que comecei a estudar Análise de Dados, mas meu conhecimento na época ainda era básico e um tanto difuso. Agora, com um conhecimento mais robusto de estatística e com muito mais segurança no tratamento de dados e criação de modelos de Machine Learning, resolvi escrever um código bem acabado e que resulta em um modelo com boa performance, além de documentá-lo da melhor maneira possível, na minha visão. Sobre a documentação, resolvi seguir a linha de trabalho do meu primeiro projeto de análise de dados, o “google-projeto-final-bicicletas” (também disponível aqui no GitHub), isto é, decidi explicar o código à medida que vou escrevendo-o com uma linguagem

amigável, de modo que uma pessoa com pouco conhecimento de Machine Learning consiga entender o que está sendo feito. Todavia, por ser um projeto um pouco mais complexo, algum conhecimento prévio é recomendado, já que eventualmente alguns termos técnicos são inevitáveis.

Métricas:

O desafio utiliza como medida de avaliação o erro quadrático médio (root-mean squared error ou RMSE, em inglês), uma medida muito comum para a avaliação de modelos de regressão. *O RMSE nesse caso específico é calculado utilizando o logaritmo do valor previsto e o logaritmo do valor real*. Utilizar o logaritmo faz com que o erro da previsão de uma casa cara afete o score do modelo da mesma maneira que o erro de uma casa barata, resultando em uma medida de avaliação mais eficaz.

O que podemos considerar um bom resultado?

A maneira mais simples de avaliar o modelo é comparar seu resultado com o dos outros usuários do Kaggle (lembrando que previsões com mais de 2 meses somem do leaderboard). Todavia, precisamos nos atentar ao fato de que algumas dessas submissões utilizam mais dados do que os disponibilizados pela competição, ou seja, estão trapaceando. Isso não gera nenhum dano já que essa competição não tem premiação, mas colocará o nosso modelo em uma colocação menor do que a real. Em resumo, podemos utilizar a análise do famoso usuário do Kaggle "fedesoriano" que concluiu que "*getting a result better than 0.10160 is very difficult without data leakage*". Essa análise pode ser encontrada neste link, lembrando que quanto menor o resultado, melhor o modelo. Com isso em mente, o melhor resultado sem trapaça dos últimos 2 meses (no momento em que escrevo essa introdução) foi de 0.11230. Ainda de acordo com a análise de *fedesoriano*, a mediana dos scores dos modelos submetidos foi de 0.1446. **Ou seja, um resultado legítimo com um score entre 0.1446 e 0.11230 pode ser considerado um bom modelo, e foi o meu score alvo. Considerando que nosso melhor score foi de 0.12365, obtivemos sucesso.** Apenas a título de curiosidade, há uma outra competição do Kaggle que utiliza os mesmos dados, porém usa como métrica apenas o RMSE, sem o logaritmo dos valores. Nesse caso o nosso melhor modelo apresentou um erro médio de 14065 dólares, um resultado excelente, dado que os melhores modelos legítimos dessa competição apresentam um erro médio de 11600 dólares.

Próximos passos:

Há diversas alterações que podemos fazer para buscar um resultado ainda melhor do nosso modelo. Algumas delas incluem: testar outros algoritmos; fazer otimização de hiperparâmetros, engenharia de atributos, seleção de atributos, fazer o pré-processamento dos dados de maneira diferente, etc. A minha ideia é atualizar esse projeto com regularidade a fim de melhorar o resultado do modelo. A próxima meta é atingir um resultado igual ou inferior à 0.11999.

Créditos e agradecimentos:

Ainda que a área de programação e TI tenha muito apreço pela ideia de código aberto (open source), fazemos justiça e exercitamos a humildade quando agradecemos àqueles que nos ensinaram. Para esse projeto, três fontes foram essenciais: o notebook "Comprehensive data exploration with Python" escrito por Pedro Marcelino e disponibilizado no Kaggle, o livro *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems, third edition*, escrito por Aurélien Geron e a fantástica videoaula/livestream "House Price Regression LIVESTREAM" produzida por Gabriel Atkin no Youtube. Meus sinceros agradecimentos às pessoas que dedicam seu precioso tempo a ensinar aos outros, muitos vezes sem ganhar um centavo com isso. Sem vocês, a jornada seria muito mais difícil.