**Overview**

**Descrição**

Comece aqui se ...

Você é novo em ciência de dados e aprendizado de máquina, ou em busca de uma simples introdução às competições de previsão de Kaggle.

Descrição da Competição

O naufrágio do RMS Titanic é um dos mais infames naufrágios da história. Em 15 de abril de 1912, durante sua viagem inaugural, o Titanic afundou depois de colidir com um iceberg, matando 1502 de 2224 passageiros e tripulantes. Esta tragédia sensacional chocou a comunidade internacional e levou a melhores normas de segurança para os navios.

Uma das razões pelas quais o naufrágio causou tal perda de vida foi que não havia botes salva-vidas suficientes para os passageiros e a tripulação. Embora houvesse algum elemento de sorte envolvido na sobrevivência do naufrágio, alguns grupos de pessoas tinham maior probabilidade de sobreviver do que outros, como mulheres, crianças e a classe alta.

Neste desafio, pedimos que você conclua a análise de quais tipos de pessoas provavelmente sobreviveriam. Em particular, pedimos que você aplique as ferramentas de aprendizado de máquina para prever quais passageiros sobreviveram à tragédia.

Praticar Habilidades

Classificação binária

Noções básicas sobre Python e R

**Avaliação**

Objetivo

É seu trabalho prever se um passageiro sobreviveu ao naufrágio do Titanic ou não.

Para cada PassengerId no conjunto de testes, você deve prever um valor 0 ou 1 para a variável Sobrevivida.

Métrica

Sua pontuação é a porcentagem de passageiros que você prevê corretamente. Isto é conhecido simplesmente como "precisão".

Formato de Arquivo de Submissão

Você deve enviar um arquivo csv com exatamente 418 entradas mais uma linha de cabeçalho. Seu envio mostrará um erro se você tiver colunas extras (além de PassengerId e Sobrevivido) ou linhas.

O arquivo deve ter exatamente 2 colunas:

PassengerId (classificado em qualquer ordem)

Sobreviveu (contém suas previsões binárias: 1 para sobrevivente, 0 para morto)

PassengerId, sobreviveu

  892,0

  893,1

  894,0

  Etc.

Você pode fazer o download de um arquivo de envio de exemplo (gender\_submission.csv) na página Dados.

**Tutoriais**

Kaggle Aprenda

O Kaggle Learn oferece cursos práticos para a maioria dos tópicos de ciência de dados. Estes cursos rápidos prepararão você rapidamente com as principais idéias para construir seus próprios projetos.

O Curso de Aprendizado de Máquina lhe dará tudo o que você precisa para ter sucesso na competição do Titanic e outros como ele.

Outros tutoriais usando Python

Comece com esta abordagem fácil de seguir para usar módulos populares do Python:

Caderno de Python Titanic Data Science Solutions

Use pandas para manipulação de dados

Use matplotlib e seaborn para visualização de dados

Aprenda a construir modelos com o scikit-learn

Esses kernels Python abrangem técnicas mais avançadas e abordagens complexas:

Um tutorial interativo sobre ciência de dados

Familiarize-se com o uso de notebooks Jupyter

Aprenda a importância da seleção de recursos no aprendizado de máquina

Aprendizado de Máquina do início ao fim com o Scikit-Learn

Use a validação cruzada para garantir que seu modelo seja generalizado para novos dados (ou seja, não "sobrecarregue")

Use o ajuste de parâmetros e a pesquisa de grade para selecionar o modelo de melhor desempenho dentre vários algoritmos de classificação diferentes

Exemplo XGBoost

Aprenda como o extremamente popular algoritmo XGBoost

Clique em "mostrar mais" na guia de código para estudar o script

Uma Introdução ao Ensemble / Empilhamento em Python:

Use a habilidade fundamental de "agrupar" para combinar as previsões de vários modelos

Outros tutoriais usando R

Explorando a Sobrevivência no Titanic

Noções básicas de engenharia de recursos e visualização de dados

Como lidar com valores ausentes no conjunto de dados

Como treinar um classificador de floresta aleatório para fazer uma previsão

Se você não está familiarizado com o Rmarkdown, clique na guia "Código" para ver o código subjacente

Famílias não são boas para a sobrevivência

Aprenda técnicas para entender como seus modelos estão fazendo previsões

Use uma visualização de um algoritmo de árvore de decisão para comparar diferentes modelos

Determinar como os recursos contribuem para a precisão da previsão

Tutoriais usando o Excel

Excel: Um tutorial sobre conceitos básicos de aprendizado de máquina em uma ferramenta familiar.

**Perguntas Frequentes**

O que é uma competição de Primeiros Passos?

As competições de Primeiros Passos foram criadas por cientistas de dados da Kaggle para pessoas que têm pouco ou nenhum histórico de aprendizado de máquina. Eles são um ótimo lugar para começar, se você é novo em ciência de dados ou acabou de terminar um MOOC e quer se envolver em Kaggle.

As competições de Primeiros Passos são uma forma não competitiva de se familiarizar com a plataforma da Kaggle, aprender conceitos básicos de aprendizado de máquina e começar a conhecer pessoas na comunidade. Eles não têm prêmio em dinheiro e estão em um cronograma contínuo.

Qual é a diferença entre um placar privado e público?

O quadro de líderes do Kaggle tem um componente público e privado para evitar que os participantes “superfiquem” o placar. Se o seu modelo for "overfit" para um conjunto de dados, ele não será generalizável fora do conjunto de dados em que você o treinou. Isso significa que seu modelo teria baixa precisão em outra amostra de dados obtida de um conjunto de dados semelhante.

Cabeçalho público

Para todos os participantes, os mesmos 50% das previsões do conjunto de testes são atribuídos ao placar público. A pontuação que você vê no placar público reflete a precisão do seu modelo nesta parte do conjunto de testes.

Classificação Privada

Os outros 50% das previsões do conjunto de testes são atribuídos ao placar privado. A tabela de classificação privada não é visível para os participantes até que a competição seja concluída. No final de uma competição, revelaremos o placar privado para que você possa ver sua pontuação nos outros 50% dos dados do teste. As pontuações na tabela de classificação privada são usadas para determinar os vencedores da competição. As competições de Primeiros Passos são executadas em um cronograma contínuo para que o placar privado nunca seja revelado.

Como faço para criar e gerenciar uma equipe?

Quando você aceita as regras da competição, uma equipe será criada para você. Você pode convidar outras pessoas para sua equipe, aceitar uma fusão com outra equipe e atualizar informações básicas, como o nome da equipe, indo para a página Mais <Equipe.

Nós ouvimos de muitos Kagglers que se unir é a melhor maneira de aprender novas habilidades e se divertir. Se você não tiver um colega de equipe, pergunte se alguém quer se juntar ao fórum de discussão.

O que são kernels?

O Kaggle Kernels é um ambiente computacional em nuvem que permite análises reprodutíveis e colaborativas. Kernels suporta scripts em relatórios R e Python, Jupyter Notebooks e RMarkdown. Vá para a guia Kernels para ver todo o código compartilhado publicamente nesta competição. Para mais informações sobre como usar o Kernels para aprender ciência de dados, visite a guia Tutoriais.

Por que minha equipe desapareceu da tabela de classificação?

Para manter o espírito das competições iniciais, implementamos uma janela rolante de dois meses sobre os envios. Depois que um envio tiver mais de dois meses, ele será invalidado e não será mais considerado para o placar.

Se a sua equipe não tiver participações nos dois meses anteriores, a equipe também será eliminada da tabela de classificação. Isso manterá a tabela de classificação em um tamanho gerenciável, renovando-a e impedindo que os recém-chegados se percam em um mar de pontuações abandonadas.

"Eu trabalhei tanto para conseguir essa pontuação! Devolva!" Leia mais sobre a nossa decisão de implementar um cabeçalho de rolagem aqui.

Como entro em contato com o suporte?

A Kaggle não tem uma equipe de suporte dedicada. Por isso, você encontrará uma resposta mais rapidamente fazendo sua pergunta no fórum apropriado. (Para esta competição, você vai querer usar o fórum de discussão do Titanic).

O suporte só pode ajudar com problemas que estão sendo experimentados por todos os participantes. Antes de entrar em contato com o suporte, verifique o fórum de discussão para obter informações sobre o seu problema. Se você não conseguir encontrá-lo, poderá postar seu problema no fórum para que um colega participante ou um membro da equipe do Kaggle possa fornecer ajuda. Os fóruns estão cheios de informações úteis sobre os dados, métricas e diferentes abordagens. Nós encorajamos você a usar os fóruns com freqüência. Se você compartilha seu conhecimento, descobrirá que outros compartilharão muito por sua vez!

Se o seu problema persistir ou se parecer eficaz para todos os participantes, entre em contato conosco.