FIATKOSKI
Data Science

DATA SCIENCE

Predição de Renda

Leonardo Fiatkoski

FIATKOSKI
Data Science

Tópicos

- Definição e Objetivo
- Descrição das Variáveis
- Análise Exploratória
- Análise Multivariada
- Modelagem Preditiva
- Análise dos Resultados

FIATKOSKI
Data Science

PREDIÇÃO DE RENDA

Definição e Objetivo

FIATKOSKI
Data Science

DEFINIÇÃO DO PROBLEMA

O dataset foi extraído em 1994 de uma base do census dos Estados Unidos e contém diversas informações relativas a vida profissional de uma amostra, com a finalidade de verificar possíveis desigualdades salariais e de condições de trabalho entre as pessoas.

OBJETIVO

Realizar uma análise exploratória para validar a hipótese de que existe desigualdade salarial na população, além da criação de um modelo preditivo para inferir quais pessoas receberão mais de 50 mil dólares por ano.

FIATKOSKI
Data Science

PREDIÇÃO DE RENDA

Descrição das Variáveis

DESCRIÇÃO DAS VARIÁVEIS

FIATKOSKI
Data Science

Variável	Descrição	Tipo	Pré-Processamento
age	Idade	int	-
workclass	Categoria da empresa em que trabalha;	object	Removidos valores marcados como '?'.
education	Grau de educação	object	-
maritalstatus	Estado civíl	object	-
occupation	Área de atuação	object	Removidos valores marcados como '?'.
relationship	Posição na família	object	Remoção dos '_' das informações de Husband e Wife.
race	Raça	object	-
sex	Sexo	object	-
capitalgain	Ganho de capital durante o ano	int	-
capitalloss	Perda de capital durante o ano	int	-
hoursperweek	Horas trabalhadas por semana	int	-
nativecountry	País de origem	object	Separação dos 3 países com maior volume e agrupamento dos demais.
over50k	Salário acima ou abaixo de 50k ano (target)	object	Binarização no momento da modelagem preditiva.

Poucas operações de pré-processamento necessárias. Não localizei nenhuma variável missing, apenas variáveis com valores deturbados (que foram removidos), ajustes no label e remoção de espaços em branco no início e fim da linha.

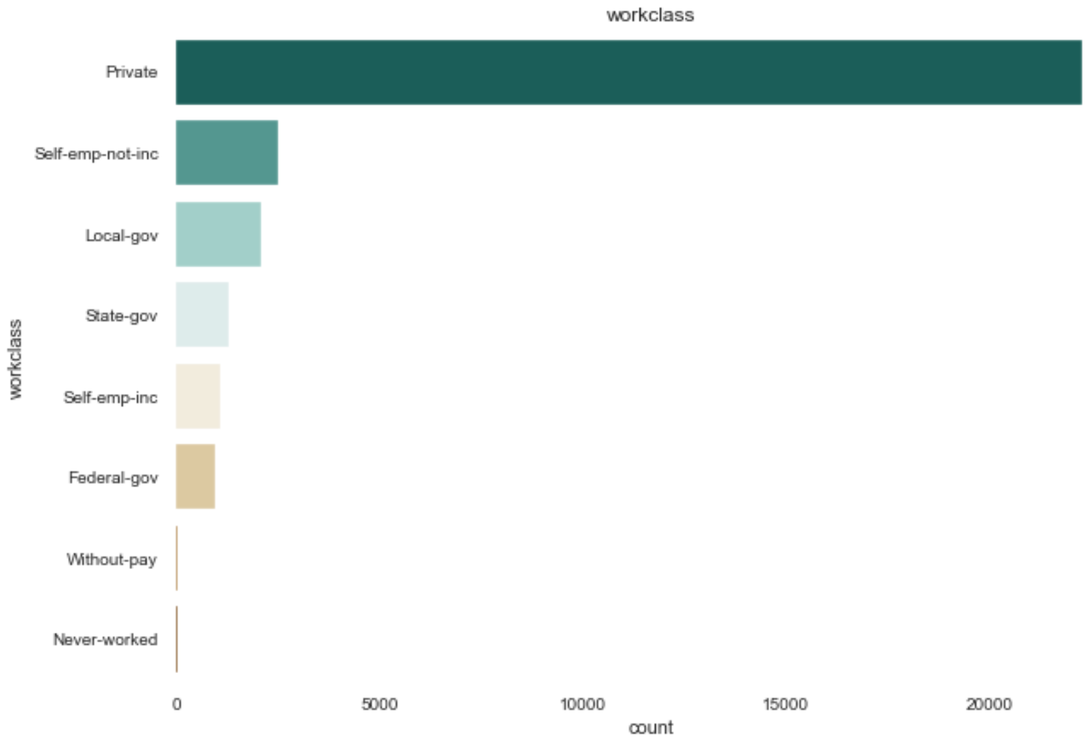
FIATKOSKI
Data Science

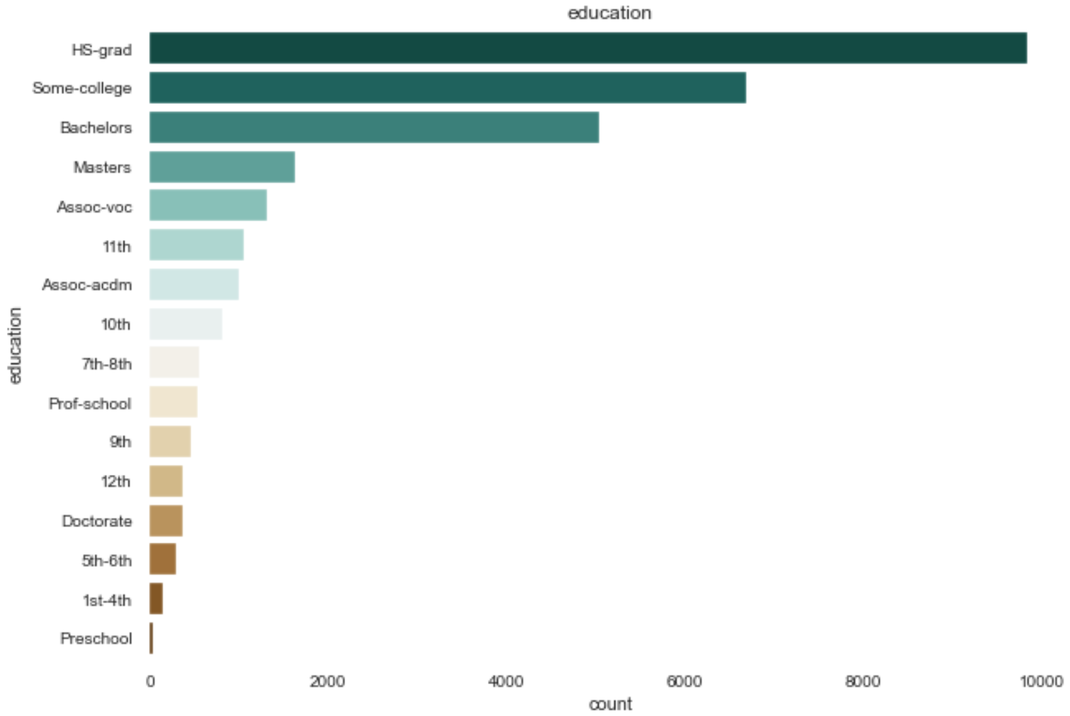
PREDIÇÃO DE RENDA

Análise Exploratória

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



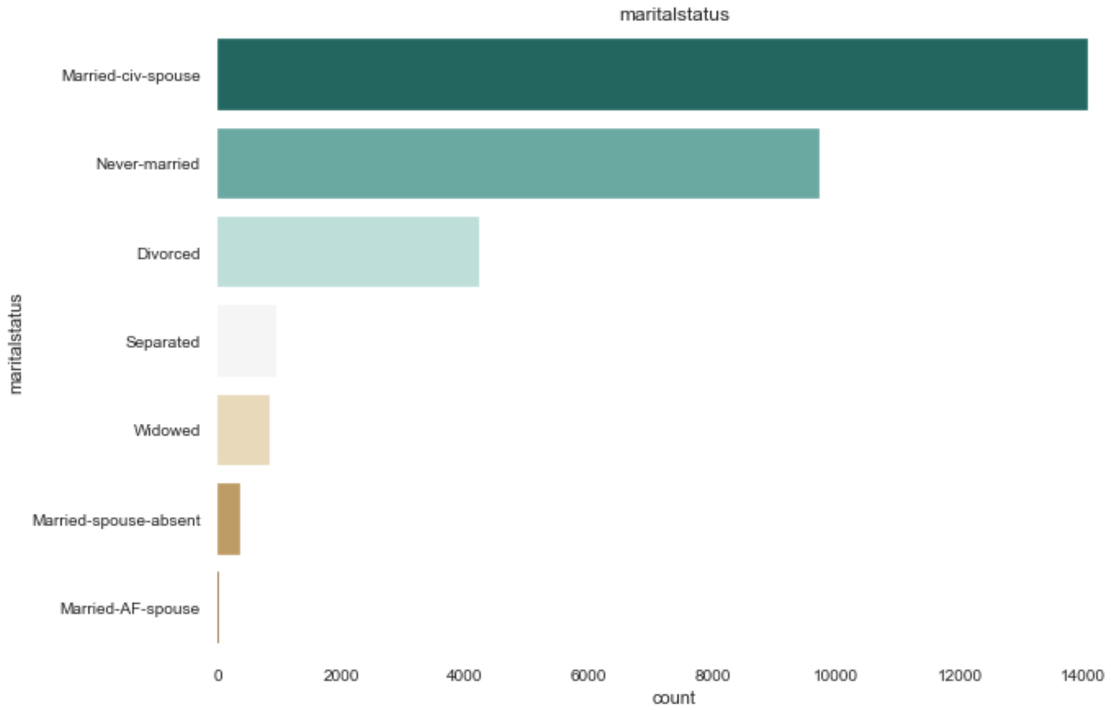


74% da base trabalha em **instituições privadas**, seguido por empreendedores e trabalhadores de estatais. Os declarantes que **não recebem salário ou que nunca trabalharam** somam **menos de 1%** de representatividade.

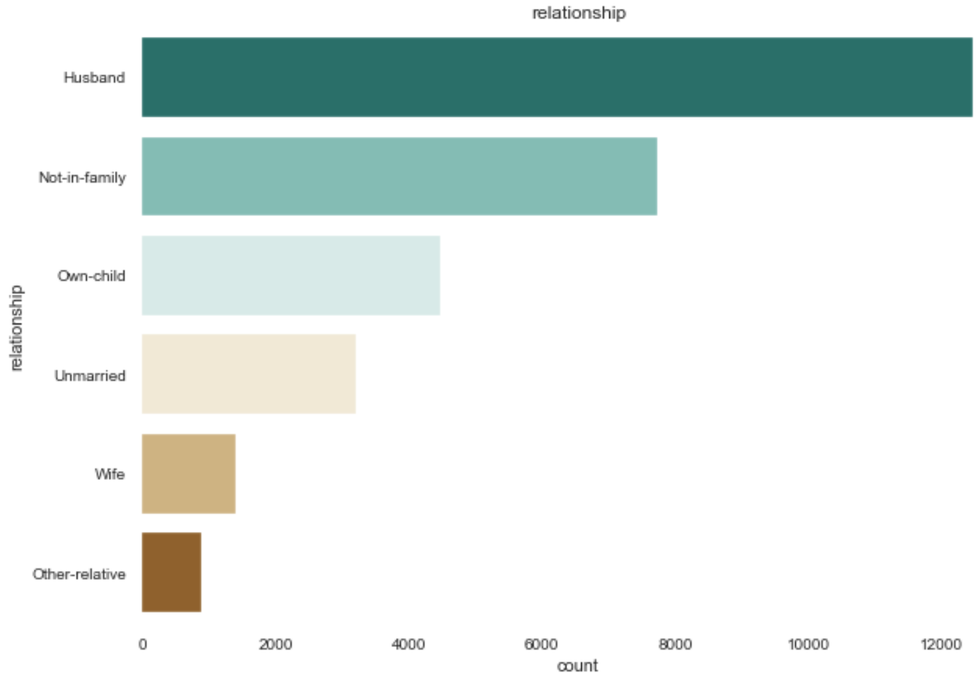
33% concluíram apenas a '**high school**', seguido com 22% dos registros como 'some college' que significa estar cursando a faculdade ou qualquer outro curso acima da high school e **17% terminaram a faculdade**.

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



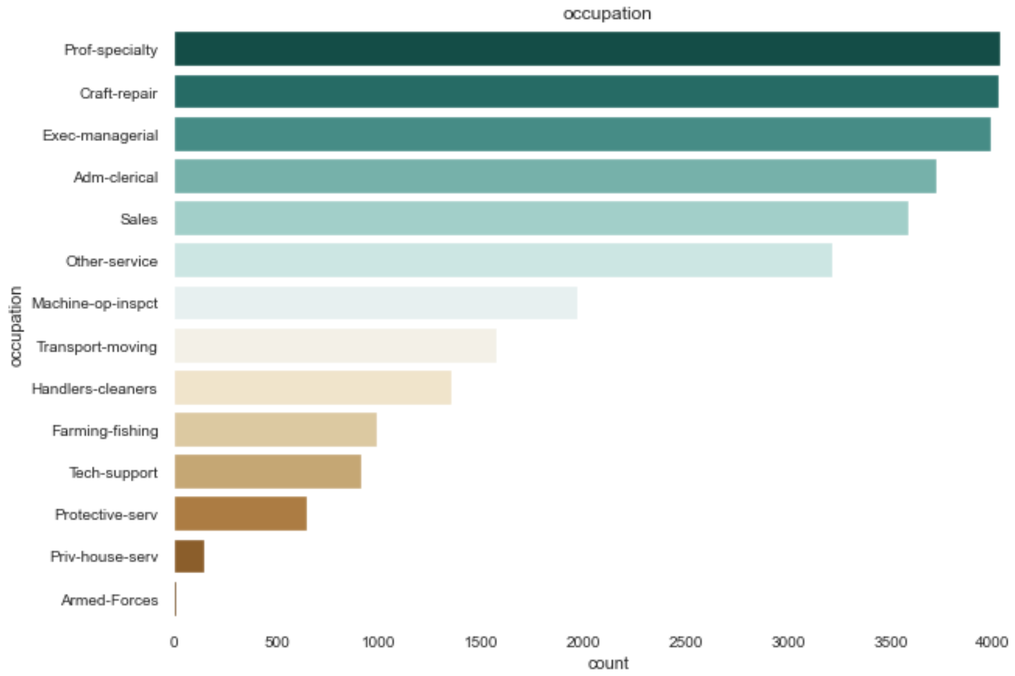
47% declararam que estão **casados**, 32% nunca casaram e 14% estão divorciados. Os 'divorciados' e 'separados' foram mantidos em variáveis diferentes por entender que se tratam de coisas diferentes. Os separados podem não ter assinado o divórcio no papel, com isso são legalmente considerados casados.

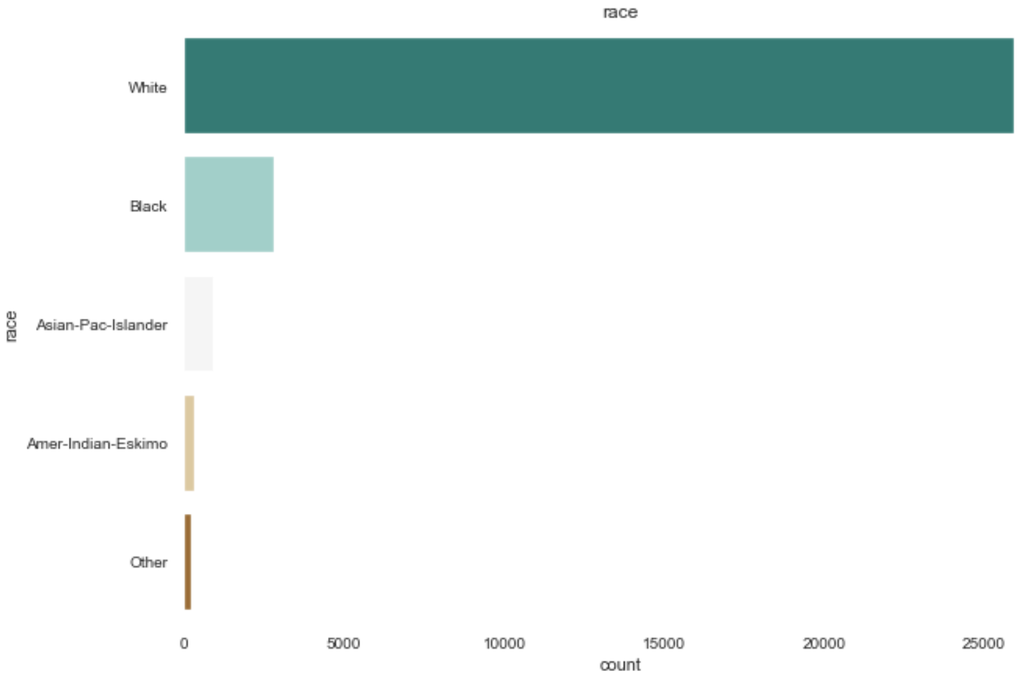


41% foram registrados como sendo o marido no status da relação (as esposas foram registradas com 6%, totalizando os 47% da variável 'maritalstatus'), seguido por 26% que declararam não pertencerem a uma família. Por se tratar de uma variável bem semelhante com a 'maritalstatus', iremos removê-la para evitar multicolinearidade.

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



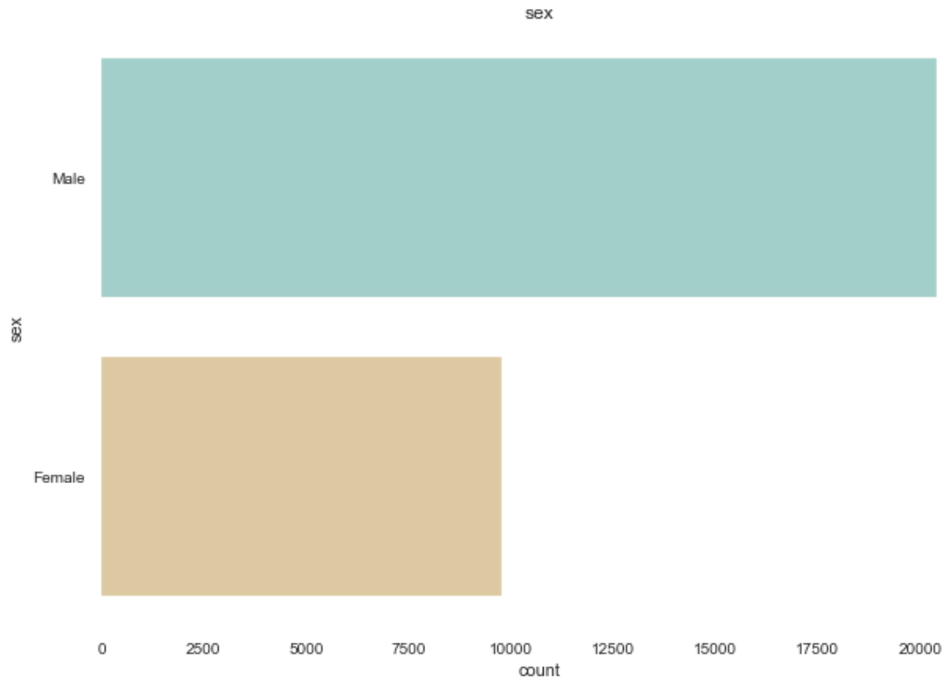


Professores Especialistas, Mecânicos e Executivos representam **13%** da base **cada**, seguidos pelos profissionais do setor administrativo e vendas com 12% casa e 11% declararam 'outros serviços', os **demais cargos** possuem **menos de 10%** de representatividade cada um.

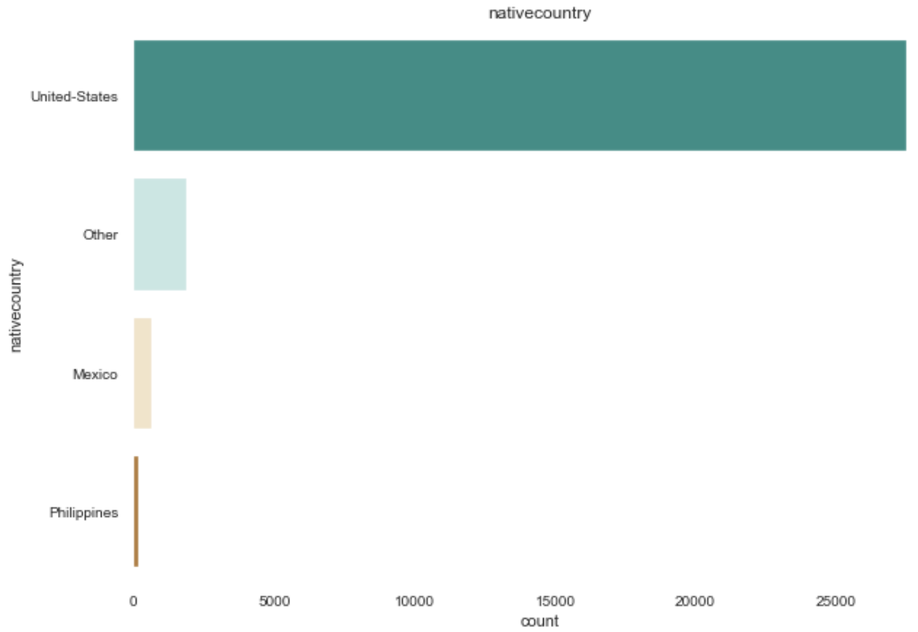
86% se consideraram de etnia branca (atualmente a representatividade no país está em 75%), contra **9% de negros**, 3% asiáticos, 1% indianos e 1% de outras etnias.

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



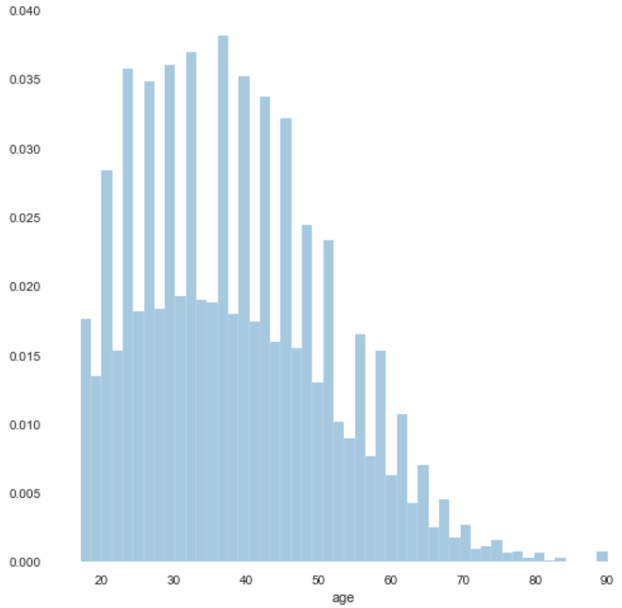
68% da base é constituída por **homens**, contra 32% de mulheres.

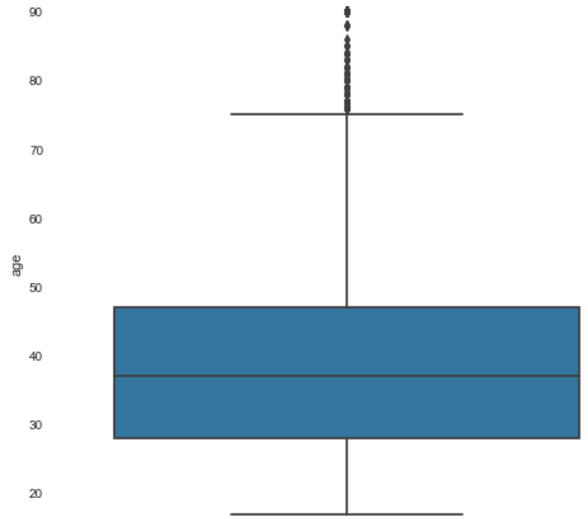


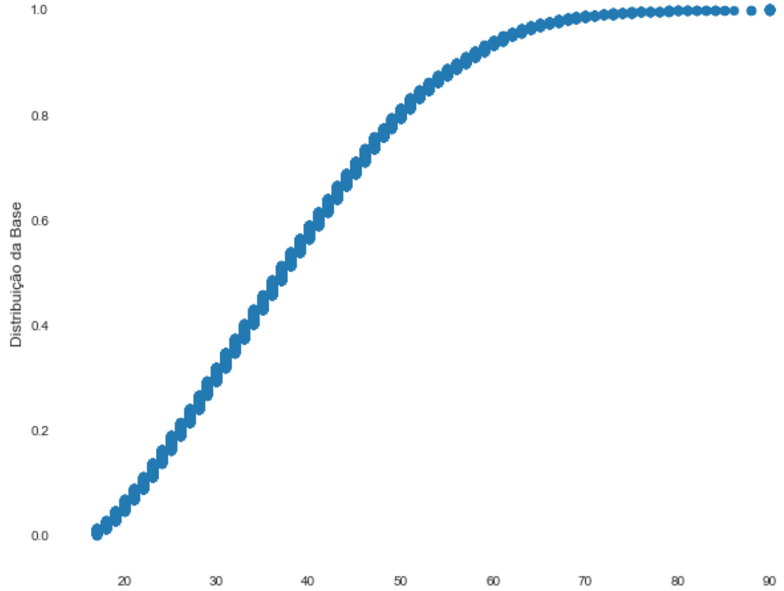
91% da base é constituída por cidadãos **americanos**, o agrupamento das diversas nacionalidades representa 6% da base, seguido de 2% de mexicanos e 1% por filipínos.

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



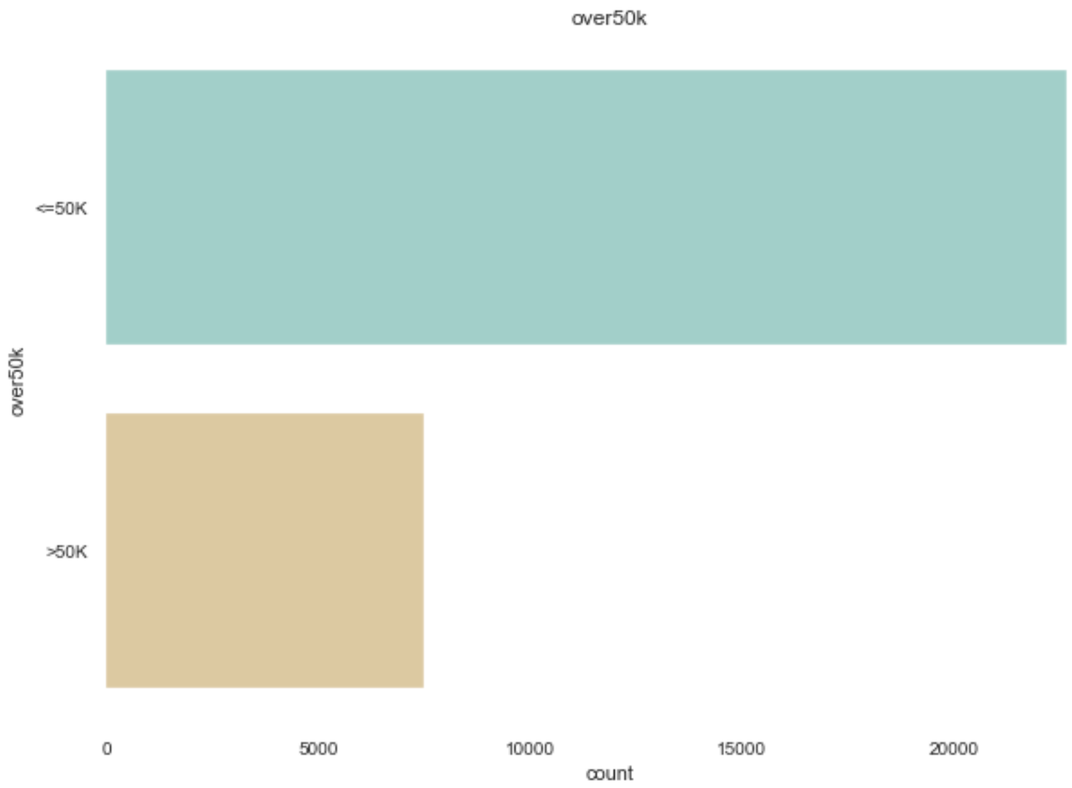




Nossa base é constituída por pessoas entre 17 e 90 anos, com **média de 38 anos** e **mediana de 37 anos**. Podemos notar um aumento no número de trabalhadores entre 20 e 30 anos, seguido por uma queda mais acentuada a partir dos 35~40 anos. Alguns outliers foram detectados acima dos 75 anos, mas serão mantidos por fazerem parte da análise em questão. Além disso, o ECDF nos mostra que cerca de **50% da distribuição** total da base possui cerca de **40 anos** de idade.

ANÁLISE EXPLORATÓRIA

FIATKOSKI
Data Science



Agora vamos para nossa variável target, se o respondente **recebe mais de 50 mil dólares por ano** ou não. Os que responderam **sim** constituem **75%** da base, contra 25% que responderam não.

obs: as classes não serão balanceadas no algoritmo, não acredito que esta proporção seja suficientemente desbalanceada para aplicação de over ou undersample.

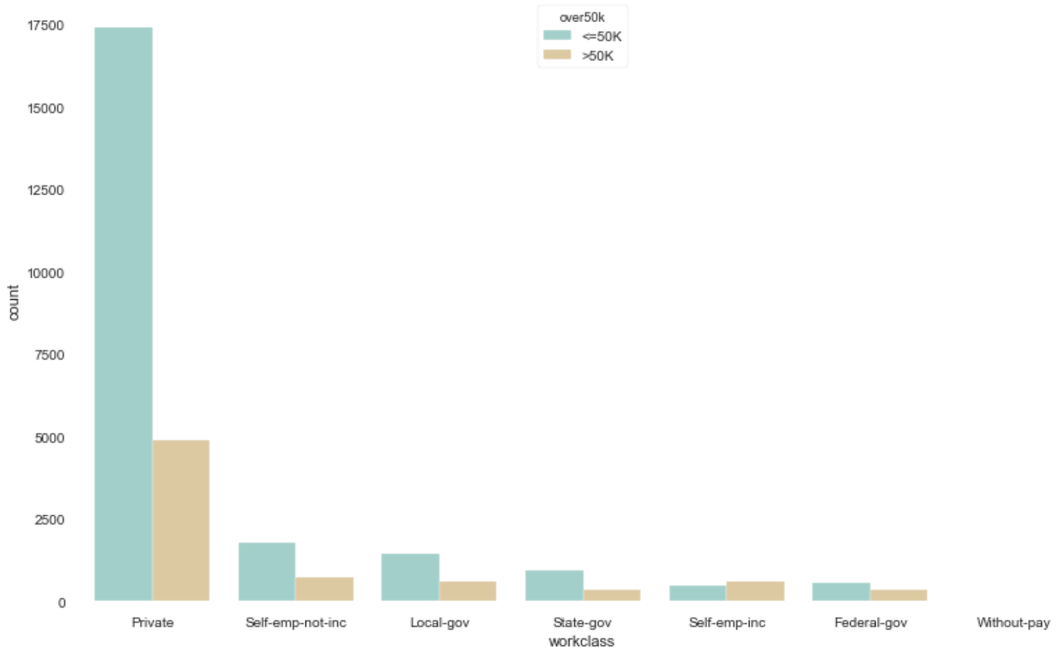
FIATKOSKI
Data Science

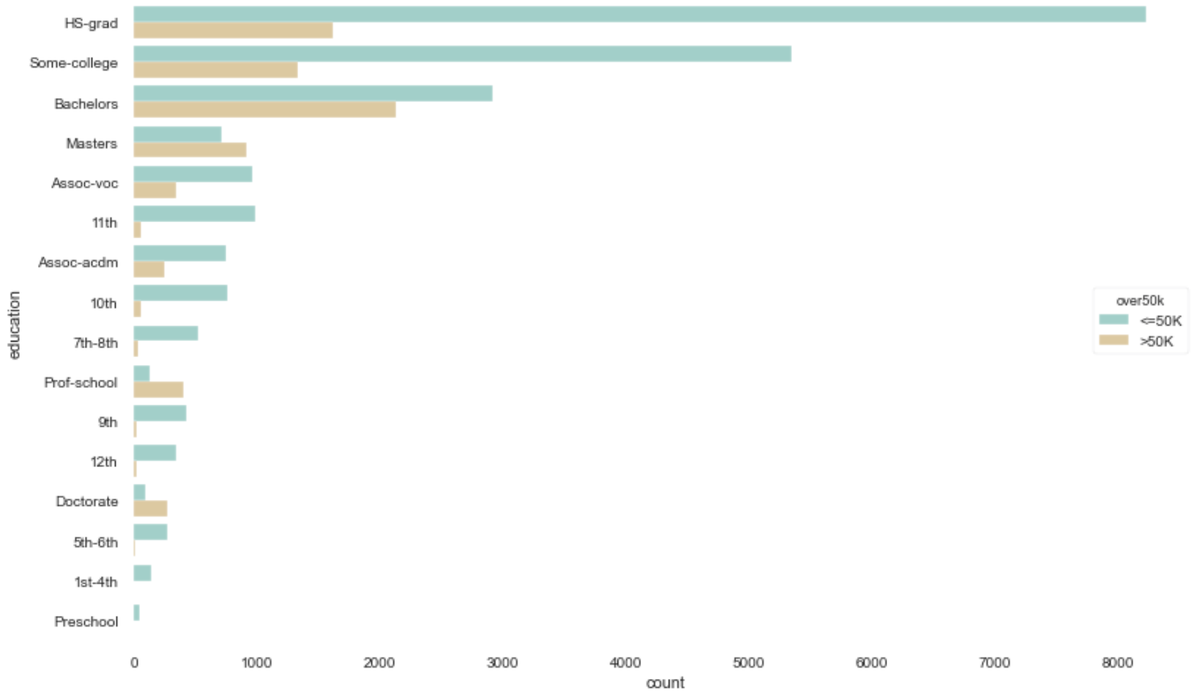
PREDIÇÃO DE RENDA

Análise Multivariada

ANÁLISE MULTIVARIADA

FIATKOSKI
Data Science



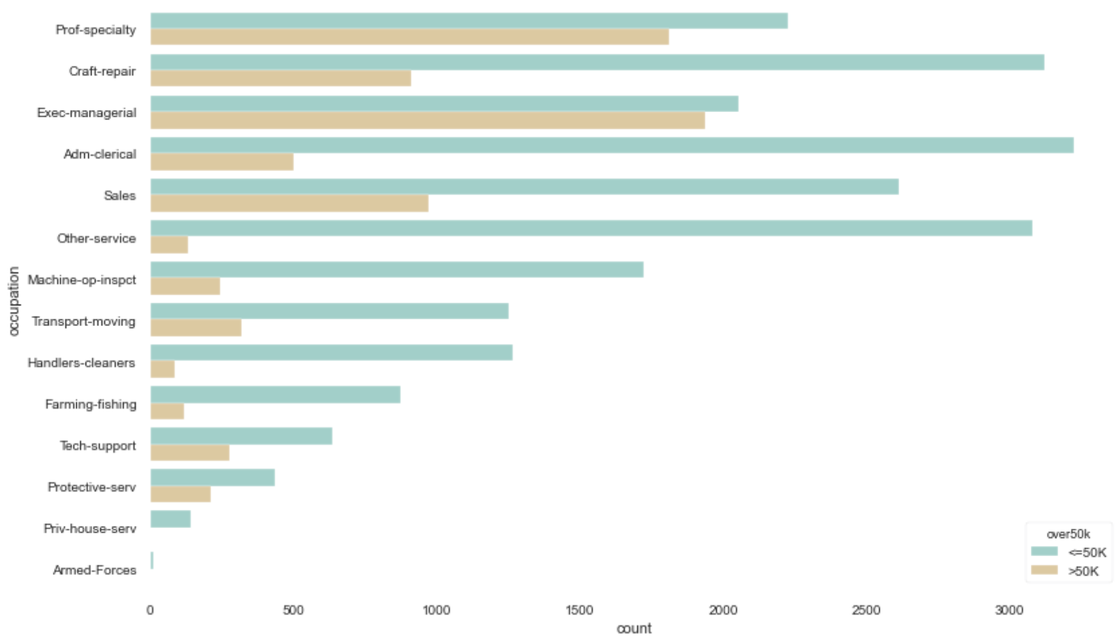


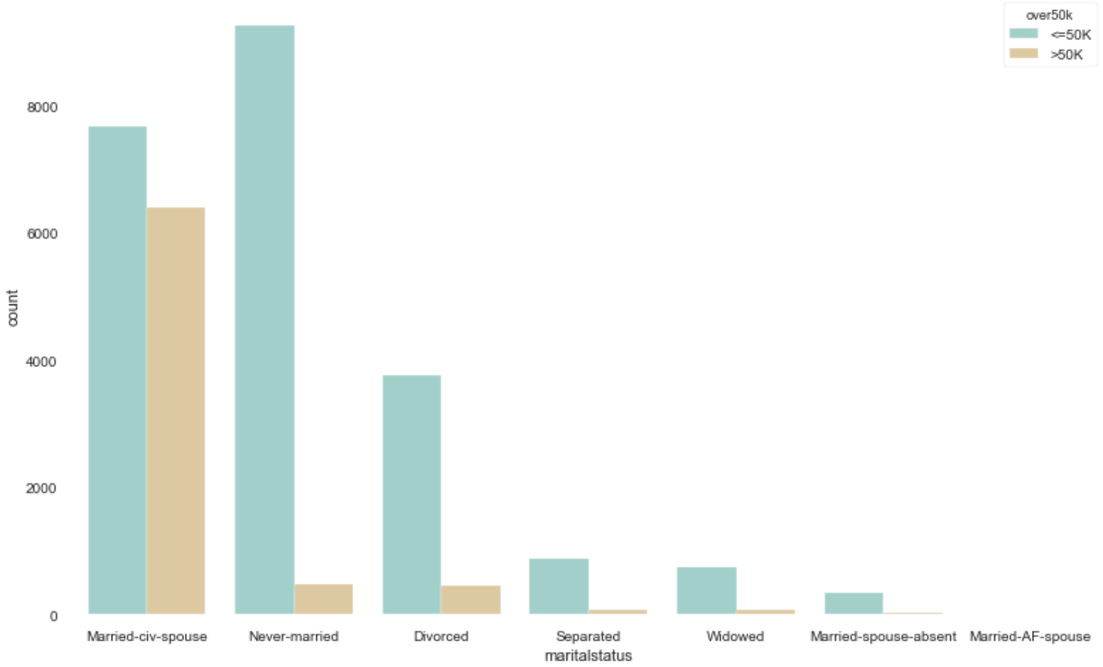
Os **empreendedores lideram** os ganhos acima de 50k com **55%** de representatividade, seguido dos **profissionais federais** com **38%**. Os funcionários de **empresas privadas** possuem o **menor índice** de ganhos com apenas **21%** ganhando acima de 50k por ano (curioso percebem que mesmo com este baixo índice, as instituições privadas ainda abrigam 74% dos profissionais)

O grau de educação com maior índice de ganhos é o de '**professional school**' com **74% ganhando acima de 50k**, que é uma modalidade de graduação para formar profissionais especialistas em determinado tema (conversa com a variável de ocupação profissional), seguido por Doutores, Mestres e Bacharelados. As pessoas que concluíram **até o 'ensino médio'** possuem um índice de ganhos acima de 50k de **menos de 10% cada**.

ANÁLISE MULTIVARIADA

FIATKOSKI
Data Science



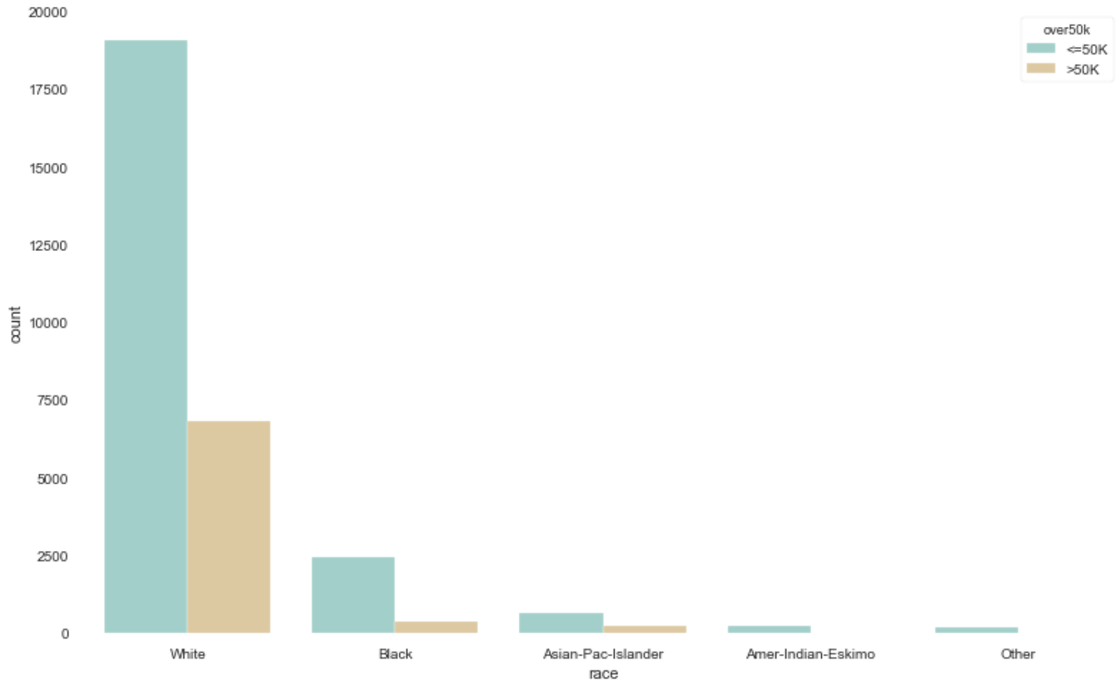


As profissões com **maior percentual de ganhos** acima de 50k é a de **Executivo** com **48%**, seguido por **Professores Especialistas** com **44%**, porém, apesar da representatividade ser maior, ainda são **classes impuras** por **não serem homogêneas**. Empregada doméstica, Outros Serviços, Limpadoras e Forças Armadas são as profissões com o **menor índice de ganhos** acima de 50k, todas representam **1% ou meno**s e possuem uma grande homogeneidade, negativa.

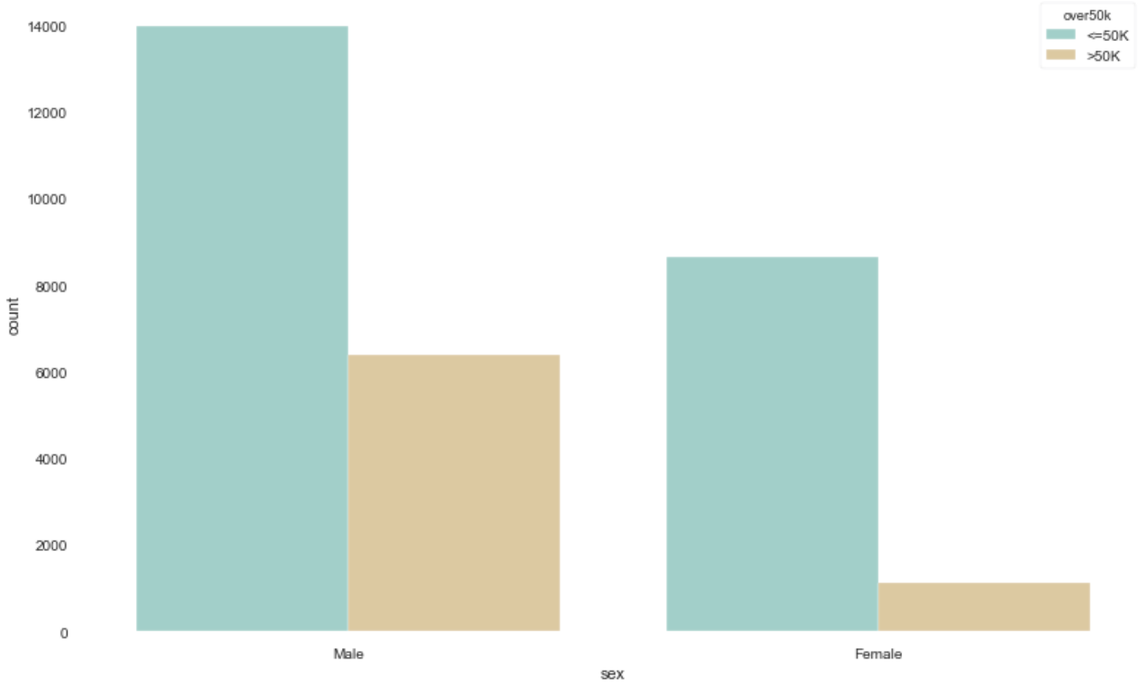
Os **casados** lideram o ranking de ganhos acima de 50k com mais de **45% de representatividade** (casais inteligentes enriquecem juntos), todos os **outros estados civis** possuem **menos de 10%** ganhando acima de 50k.

ANÁLISE MULTIVARIADA

FIATKOSKI
Data Science



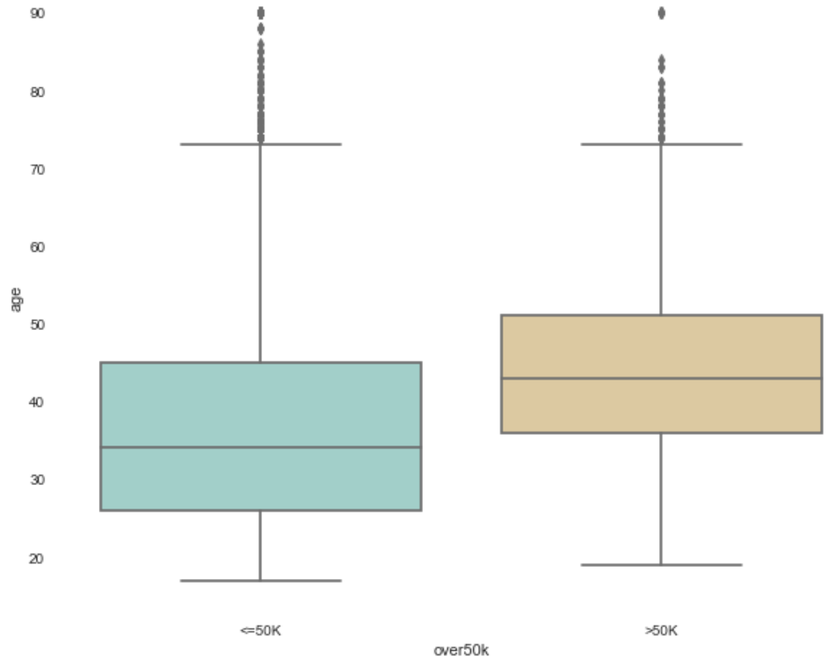
Falando em etnia, os **asiáticos** possuem a **maior representatividade nos altos ganhos** com cerca de **27%**, seguido por brancos com 26%. Negros e Indianos possuem 12% e 11% respectivamente, seguidos por outros com 9%.



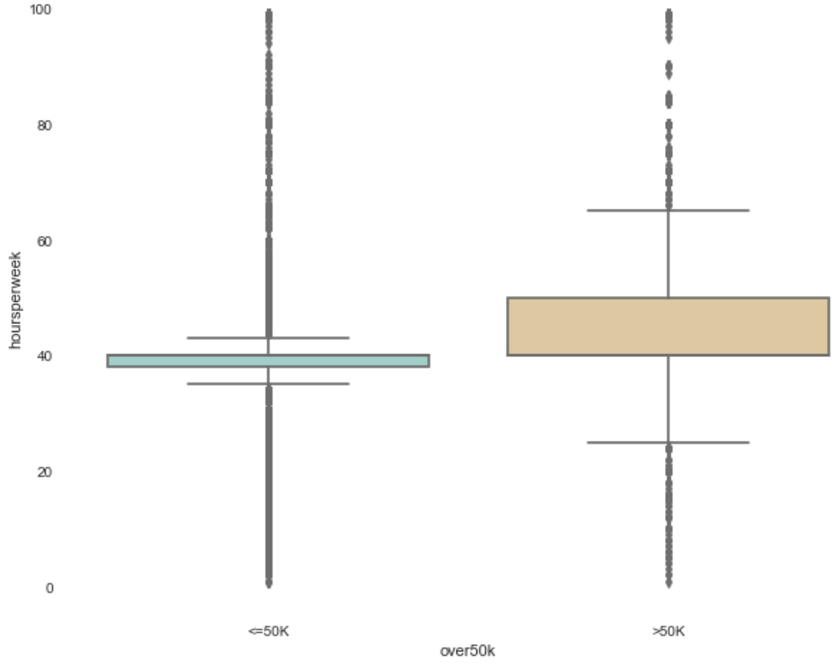
Na variável sexo, os **homens** com ganhos acima de 50k são **31%**, contra **apenas 12% das mulheres** (pois é...)

ANÁLISE MULTIVARIADA

FIATKOSKI
Data Science



Em relação a idade, percebemos uma **forte relação** da idade **com os ganhos**, a **média de idade** de quem ganha **menos** de 50k por ano é de **33 anos**, já a **média** de idade de quem ganha **mais de 50k** é de **44 anos**.



O gráfico nos indica que aqueles com maior média de horas trabalhadas possuem uma maior relação com os ganhos acima de 50k.

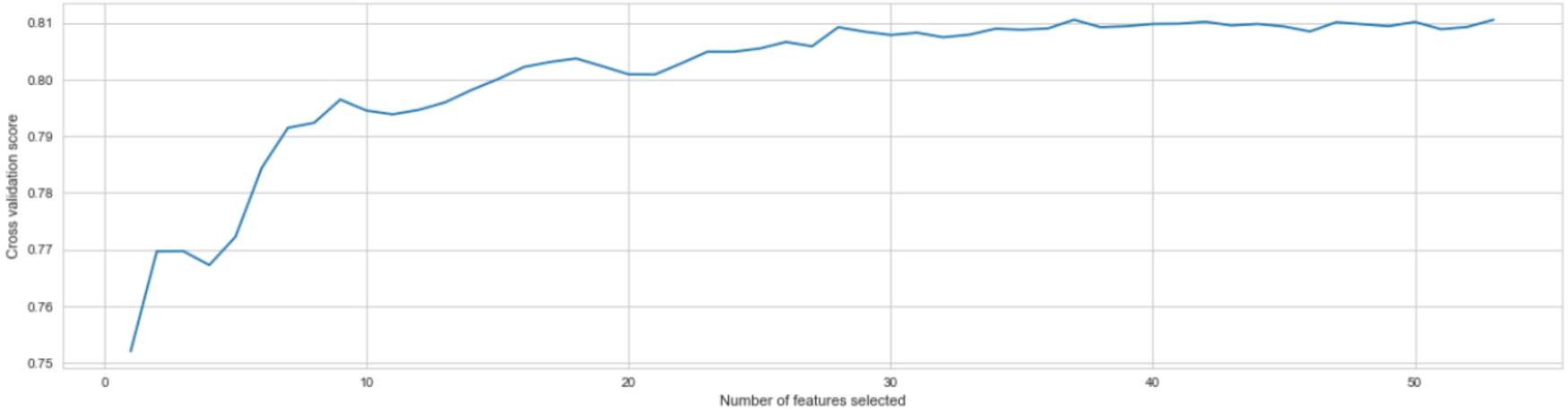
FIATKOSKI
Data Science

PREDIÇÃO DE RENDA

Modelagem Preditiva

SELEÇÃO DE FEATURES

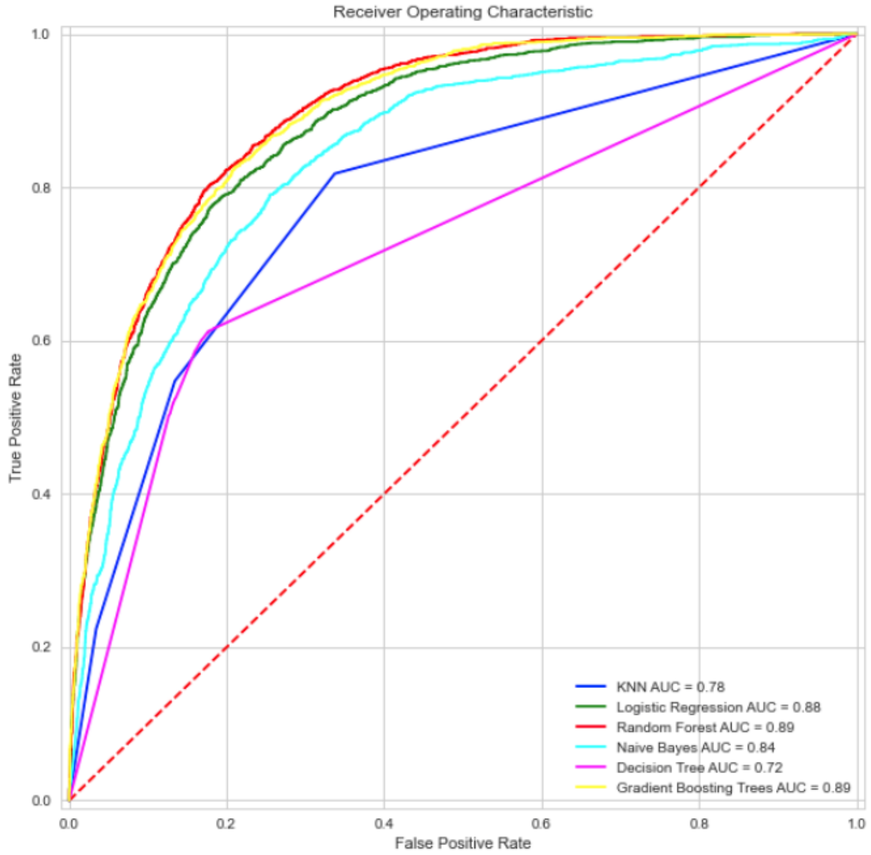
FIATKOSKI
Data Science



Para realizar a seleção das features, utilizamos uma técnica chamada RFE (Recursive Feature Elimiation) que basicamente é uma seleção retroativa dos preditores. Essa técnica começa construindo um modelo em todo o conjunto de preditores e calcula uma pontuação de importância para cada preditor. Os preditores menos importantes são removidos, o modelo é reconstruído e as pontuações de importância são computadas novamente. Portanto, o tamanho do subconjunto é o parâmetro de ajuste para o RFE e otimiza os critérios de desempenho usados para selecionar os preditores com base nas classificações de importância.

SELEÇÃO DO MODELO

FIATKOSKI
Data Science



Para avaliação do modelo utilizamos a curva ROC, que nos mostra o quão bom o modelo criado pode distinguir entre duas classes.

O ROC possui dois parâmetros:

- Taxa de verdadeiro positivo (True Positive Rate), que é dado por true positives / (true positives + false negatives)
- Taxa de falso positivo (False Positive Rate), que é dado por false positives / (false positives + true negatives)

Uma curva ROC traça “True Positive Rate vs. False Positive Rate” em diferentes limiares de classificação (figura ao lado).

Assim, na tentativa de simplificar a análise da ROC, a AUC (“area under the ROC curve”) nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da ROC, calculando a “área sob a curva”.

O modelo selecionado foi o **Gradient Boosting Classifier**, com **89% de AUC** no modelo base e **84% utilizando Cross-Validation** de 10 folds.

RESULTADOS FINAIS

FIATKOSKI
Data Science

	precision	recall	f1-score	support
0	0.86	0.94	0.90	5642
1	0.75	0.55	0.64	1899
accuracy			0.84	7541
macro avg	0.81	0.75	0.77	7541
weighted avg	0.83	0.84	0.83	7541

Precision: Verdadeiro Positivo / Verdadeiros Positivos + Falsos Positivos
Daqueles que classifiquei como corretos, quantos efetivamente eram?

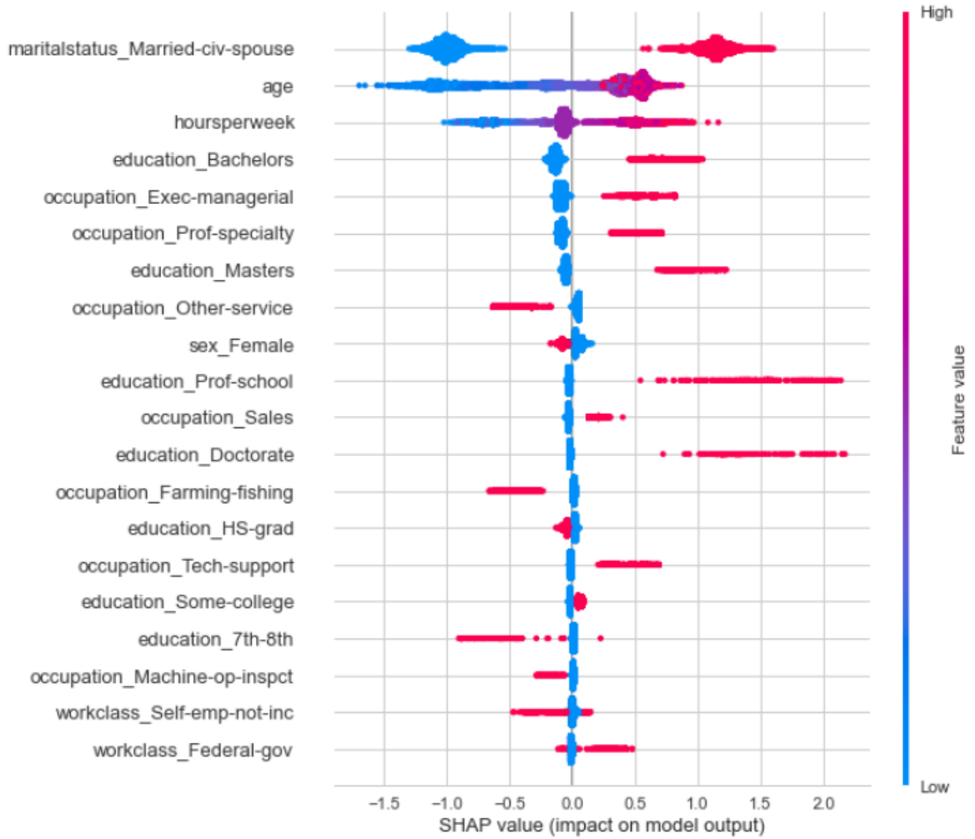
Recall: Verdadeiro Positivo / Verdadeiros Positivos + Falsos Negativo
Quando realmente é da classe X, o quão frequente classifiquei como X?

F1: 2 * Precision * Recall / Precision + Recall
Essa métrica combina Precision e Recall, de modo a trazer um número único que indique a qualidade geral do seu modelo

Podemos verificar que conseguimos predizer com maior precisão a classe 0 (recebe menos de 50 mil dólares ao ano) com 90% de F1-Score, enquanto a classe 1 possui 64%.
No geral são bons resultados, ainda mais considerando que os dados possuem um certo desbalanceamento, no qual optamos por não realizar tratamentos de over ou undersample.

SHAP

FIATKOSKI
Data Science



Com o SHAP podemos ver quanto cada variável contribuiu para o resultado final da predição:
- Quanto maior o valor da variável, maior a intensidade da cor vermelha;
- Quanto mais pra direita o dado se encontra, maior a força da predição para a classe acima de 50k.

obs: em caso de variável categórica, 1 é vermelho e 0 é azul.

E como se era esperado por conta da análise multivariada, as variáveis com maior contribuição no modelo foram o status de casado, idades altas, maior média de horas trabalhadas, educação a partir de bacharelado e cargos de executivos e professores especialistas.

DATA SCIENCE

Obrigado.

Leonardo Fiatkoski