Relatório Técnico

Descrição do problema

A Khipo busca analisar e entender o sistema de precificação dos apps de motoristas para poder, com o poder da inteligência artificial, fazer um predição do preço de uma determinada corrida baseado nas condições apresentadas de antemão, como: clima, endereço de origem e destino, horário da corrida, trânsito.

Tratamento do Dataset

Ferramentas Utilizadas

- Numpy
- Pandas

Organizando os Títulos das Colunas

O primeiro passo tomado no tratamento do Dataset fornecido foi organizar as colunas, traduzindo o nome de cada uma para português, padronizando o nome das colunas e organizando elas em uma melhor disposição, este passo foi realizado para facilitar a detecção de cada dado específico e para melhorar a legibilidade geral das tabelas.

Tabela Dados de Produto:

Antes

```
ProductID  ProviderID  CategoryID        Description
```

Depois

```
ID_Produto  ID_Provedor  ID_Categoria        Descricao
```

Tabela Dados da Corrida:

Antes

RideID	UserID	Schedule	Create	RideStatusID	CompanyID	ProviderID	RideProviderID

price	Updated	CategoryID	TotalUsers	Car	RideDriverLocationID	ScheduledRide

Depois

ID_Corrida	ID_Usuario	Agendamento	Criado	ID_Status_Corrida	ID_Empresa	ID_Provedor	ID_Provedor_Corrida

Preco	Atualizado	ID_Categoria	Usuarios_Total	Carro	ID_Local_Motorista	Corrida_Agendada

Tabela Dados do Endereço:

Antes

RideAddressID	Address	Street	Number	Neighborhood	City	State	Lat	Lng	RideAddressTypeID	RideID

Depois

ID_Endereco_Corrida	Endereco	Rua	Numero	Bairro	Cidade	Estado	Latitude	Longitude	ID_Tipo_Endereco_Corrida	ID_Corrida

Tabela de Estimativa da Corrida:

Antes

RideEstimativeID	RideID	ProductID	WaitingTime	Price	FareID	Selected	RideReasonSelectedEstimativeID	Fee

Depois

ID_Estimativa_Corrida	ID_Corrida	ID_Produto	Tempo_Espera	Preco	ID_Tarifa	Selecionado	ID_Motivo_Selecionado_Estimativa	Taxa

Código da Renomeação de Colunas:

```
mapaProduct = {
    "ProductID" : "ID_Produto",
    "ProviderID" : "ID_Provedor",
    "CategoryID" : "ID_Categoria",
    "Description" : "Descricao",
}

mapaRide = {
    "RideID" : "ID_Corrida",
    "UserID" : "ID_Usuario",
    "Schedule" : "Agendamento",
    "Create" : "Criado",
    "RideStatusID" : "ID_Status_Corrida",
    "CompanyID" : "ID_Empresa",
```

```
    "ProviderID" : "ID_Provedor",
    "RideProviderID" : "ID_Provedor_Corrida",
    "price" : "Preco",
    "Updated" : "Atualizado",
    "CategoryID" : "ID_Categoria",
    "TotalUsers" : "Usuarios_Total",
    "Car" : "Carro",
    "RideDriverLocationID" : "ID_Local_Motorista",
    "ScheduledRide" : "Corrida_Agendada",
}

mapaRideAddress = {
    "RideAddressID" : "ID_Endereco_Corrida",
    "Address" : "Endereco",
    "Street" : "Rua",
    "Number" : "Numero",
    "Neighborhood" : "Bairro",
    "City" : "Cidade",
    "State" : "Estado",
    "Lat" : "Latitude",
    "Lng" : "Longitude",
    "RideAddressTypeID" : "ID_Tipo_Endereco_Corrida",
    "RideID" : "ID_Corrida",
}

mapaRideEst = {
    "RideEstimativeID" : "ID_Estimativa_Corrida",
    "RideID" : "ID_Corrida",
    "ProductID" : "ID_Produto",
    "WaitingTime" : "Tempo_Espera",
    "Price" : "Preco",
    "FareID" : "ID_Tarifa",
    "Selected" : "Selecionado",
    "RideReasonSelectedEstimativeID" :
"ID_Motivo_Selecionado_Estimativa",
    "Fee" : "Taxa",
}

dadosProduto.rename(columns=mapaProduct, inplace=True)
dadosRide.rename(columns=mapaRide, inplace=True)
dadosRideAddress.rename(columns=mapaRideAddress, inplace=True)
dadosRideEst.rename(columns=mapaRideEst, inplace=True)
```

Excluindo valores repetidos

A seguir detectamos que na tabela de dados de endereço uma das colunas possuía o endereço completo, e as 5 tabelas depois dela apresentavam esse mesmo valor quebrado em diferentes partes (rua, número, bairro, cidade e estado) então excluímos essas 5 tabelas deixando todos esses valores agregados em apenas uma que pode ser extraído conforme a necessidade.

Antes

Address	Street	Number	Neighborhood	City	State
Rua João Pinheiro, 585 - Rua João Pinheiro - B...	Rua João Pinheiro	585	Rua João Pinheiro	NaN	Brasil
Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...	450	NaN	NaN	NaN
Rodovia Rafael da Rocha Pires, 1883 - Rodovia ...	Rodovia Rafael da Rocha Pires	1883	Rodovia Rafael da Rocha Pires	NaN	Brasil
Angeloni Ingleses (Florianópolis) - Supermerca...	Angeloni Ingleses (Florianópolis) - Supermercado	6375	NaN	NaN	NaN
Rua Barão do Rio Branco, 12 - Rua Barão do Rio...	Rua Barão do Rio Branco	12	Rua Barão do Rio Branco	NaN	Brasil

Depois

Endereco
Rua João Pinheiro, 585 - Rua João Pinheiro - B...
Av. Dr. Nereu Ramos, 450 - Rocio Grande, São F...
Rodovia Rafael da Rocha Pires, 1883 - Rodovia ...
Angeloni Ingleses (Florianópolis) - Supermerca...
Rua Barão do Rio Branco, 12 - Rua Barão do Rio...

Código da Exclusão de colunas repetidas:

```
dadosRide.drop(columns=['Agendamento', "Atualizado",
"ID_Provedor_Corrida", "Carro", "ID_Local_Motorista", "ID_Usuario"])
dadosRideAddress.drop(columns=['Rua', "Numero", "Bairro", "Cidade",
"Estado"])
dadosRideEst.drop("ID_Motivo_Selecionado_Estimativa", axis='columns')
```

Justificativa das Exclusões

- Tabela Dados da Corrida:
    - Linhas “Agendamento” e “Atualizado - Possuem valores quase idênticos aos da coluna “Criado” com apenas alguns microssegundos de diferença
    - Linhas “Carro” e “ID_Local_Motorista” - Apresentam quase unicamente valores nulos, e as poucas exceções que possuem valores não apresentam relevância para cálculo do preço
    - Linhas “ID_Provedor_Corrida” e “ID_Usuario” - Não apresentam valores relevantes para o cálculo do preço

- Tabela Dados de Endereço
    - Linhas ”Rua”, “Numero”, “Bairro”, “Cidade”, “Estado” - Apresentam valores repetidos da coluna “Endereço”

- Tabela de Estimativa da Corrida
    - Linha “ID_Motivo_Selecionado_Estimativa” - Não apresenta valores relevantes para o cálculo do preço

Reordenando as Colunas

Então trouxemos a coluna ID_Corrida (coluna em comum entre todas as tabelas) para frente da tabela, para facilitar a identificação individual de cada linha através das diferentes tabelas e assim poder associar os dados de uma tabela com as outras.

Código da Reordenação de Colunas:

```
dadosRideAddress = dadosRideAddress[["ID_Corrida",
"ID_Endereco_Corrida", "Endereco",
                                     "Latitude", "Longitude",
"ID_Tipo_Endereco_Corrida"]]

dadosRideEst = dadosRideEst[["ID_Corrida", "ID_Produto",
"ID_Estimativa_Corrida", "Tempo_Espera",
                             "Preco", "ID_Tarifa", "Selecionado",
"Taxa",]]
```

Análise Inicial:

Uma análise inicial superficial nos ajuda a começar a entender as características que são levadas em consideração no cálculo do preço das viagens dos apps, assim como também nos permite começar a entender como identificar o peso que cada característica tem nesse cálculo que será melhor elaborado com uma análise mais profunda.