Relatório

Resumo

O objetivo deste exercício é a criação de um algoritmo de machine learning no qual seja capaz de prever o preço de uma viagem de táxi. Como neste problema não foi fornecida qualquer informação sobre o dataset em estudo, levou a que todas as decisões tomadas no tratamento e processamento da mesma tenham sido baseadas no meu conhecimento sobre o mundo. Neste exercício estou a considerar como realidade um serviço de táxis no qual existe um valor mínimo de uma viagem, existe uma distância mínima para a qual a viagem se irá realizar, existe um tempo mínimo de viagem e a velocidade média não ultrapassará certos limites e existirão passageiros no taxi.

Pré processamento

Durante o pré processamento foram excluídas todas as viagens que possuíam as seguintes propriedades:

- Custo zero
- Tempo de viagem inferior a 30 segundos
- Número de passageiros igual a zero

Criou-se também uma auxiliar chamada “speed” para a qual nos ajudou a excluir todas as viagens em que a velocidade média é superior a 100 e inferior a 2.

Algoritmo

O algoritmo escolhido para o nosso modelo foi a regressão linear. Apesar de existirem algoritmos mais robustos que iteram até obter a maior precisão e que atenuam a influência de outliers, a regressão linear revelou ser a escolha mais precisa entre os algoritmos testados.

Resultados obtidos

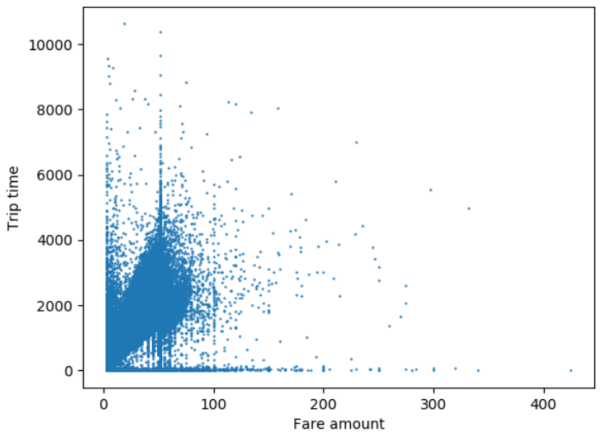
Features	Pré processamento	Variância	Média de erro quadrático
Todas	Não	0.87	11.87
rate_code trip_distance trip_time_in_secs	Não	0.87	11.93
Todas	Sim	0.95	4.21
rate_code trip_distance trip_time_in_secs	Sim	0.95	4.22

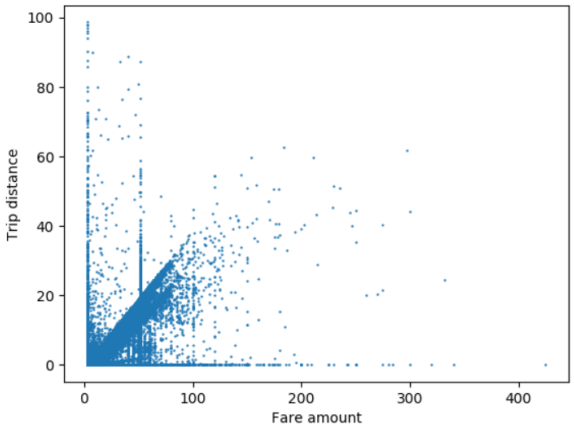
Conclusões

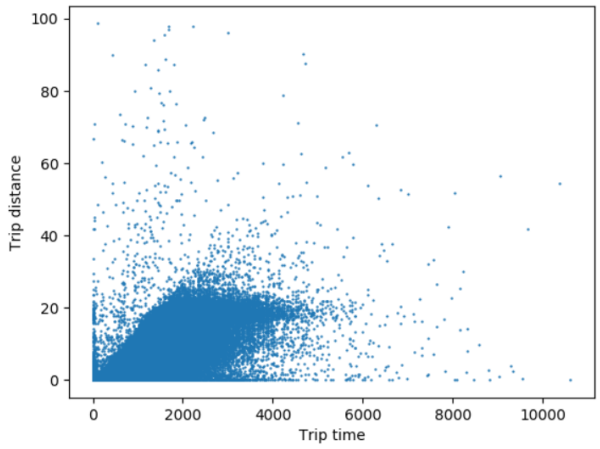
Na tabela acima podemos verificar que apesar de uma pequena diferença na média do erro quadrático, as features que influenciam diretamente o preço são rate_code, trip_distance e trip_time_in_secs. Podemos também observar que com a aplicação do pré processamento existe uma melhoria considerável na precisão do modelo criado levando a que previsões realizadas sejam mais precisas.

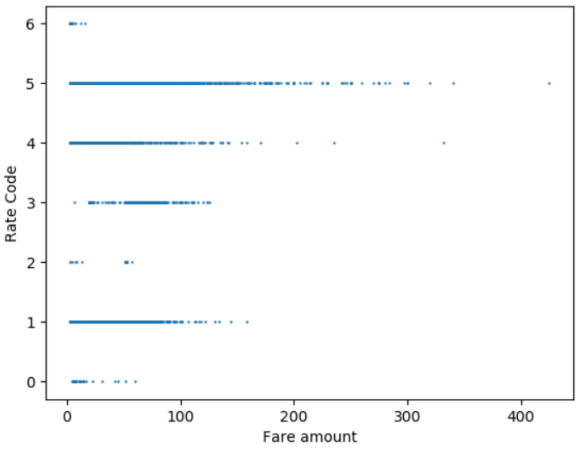
Anexos

Dados sem pré processamento









Dados com pré processamento