2º Semestre 2021/2022

Previsão no dataset Seoul Bikes

Projeto Aplicado em Ciência de Dados I

Docentes Sérgio Moro e Diana Mendes

Grupo 12

André Novo nº 93343

Beatriz Paulino nº98484

Luís Pereira nº 98398

Maria Pais nº 98263

Sebastião Rosalino nº 98437

iscte INSTITUTO UNIVERSITÁRIO DE LISBOA

Business Understanding

Introdução

Em 2015 foi fundado o serviço de alugueres de bicicletas **Seoul Bikes** que visa a promoção de um meio de locomoção sustentável para os cidadãos da Coreia do Sul. Este sistema possui estações dispersas por Seoul nas quais é possível efetuar o aluguer de bicicletas, sendo estas consideradas pontos de entrega e recolha das mesmas. Com recurso à aplicação da Seoul Bikes pode ser verificada a disponibilidade das bicicletas, algo que facilita o processo logístico associado.

Formulação do problema

A partir de um conjunto de dados públicos referentes aos alugueres de bicicletas da companhia Seoul Bikes, juntamente com as condições meteorológicas que afetam este mesmo aluguer, o **objetivo deste projeto** consiste em **prever o número de bicicletas necessárias** a disponibilizar a cada hora de forma a tornar o serviço o mais eficiente possível.

Critérios de sucesso

Perante esta situação é pertinente ter em conta os riscos associados, nomeadamente, a escassez de bicicletas disponibilizadas face à procura num determinado ponto de recolha. Como medidas preventivas contra tais riscos considerou-se que deveria ser estabelecida uma **margem mínima de bicicletas disponíveis** com o intuito de maximizar o lucro assegurando em simultâneo a disponibilidade do serviço.
Em suma, considera-se que o trabalho é bem sucedido caso seja obtido um modelo com **capacidade de generalização e um erro de previsão mínima.**

Ferramentas

Para a realização deste projeto contamos com 5 alunos de 2º ano da licenciatura de Ciência de Dados, um docente supervisor, um dataset proveniente do UCI Machine Learning Repository, computadores pessoais , Jupyter Notebook, linguagem de programação Python, Prezi e Django.

Data Understanding

O dataset a ser utilizado neste projeto contém informação relativa ao aluguer de bicicletas por hora em Seoul desde Dezembro de 2017 até Novembro de 2018. Este possui 8760 observações e 14 variáveis, em que cada registo corresponde a uma hora do dia.

Segue-se a descrição das variáveis da base de dados:

Coluna	Tipo	Descrição
Date	String	O dia do ano, durante 365 dias
Rented Bike Count	Inteiro	Número de bicicletas alugadas por hora
Hour	Inteiro	Hora do dia
Temperature(°C)	Float	Temperatura por hora
Humidity(%)	Inteiro	Humidade no ar
Wind Speed (m/s)	Float	Velocidade do vento em metros por segundo
Visibility(10m)	Inteiro	Visibilidade por 10 metros
Dew Point Temperature(°C)	Float	Temperatura no início do dia
Solar Radiation (MJ/m2)	Float	Radiação solar
Rainfall(mm)	Float	Chuva por milímetro
Snowfall(cm)	Float	Neve por centímetro
Seasons	String	Estação do ano
Holiday	String	Se é ou não feriado nacional
Functioning Day	String	Se o serviço de aluguer estava ou não em funcionamento

Em primeira instância importou-se o dataset para JupyterNotebook com recurso à biblioteca Pandas. Verificou-se o tipo de dados de cada variável e chegou-se à conclusão de que a variável Date foi importada num formato inviável **(string em vez de Datetime)**, procedendo-se à sua alteração.

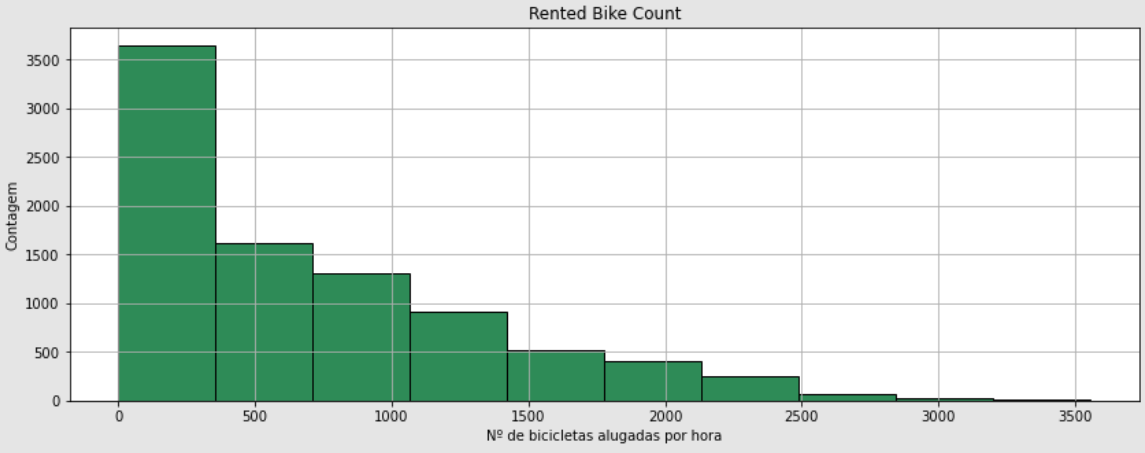
Continuando a análise exploratória dos dados, viram-se as estatísticas descritivas básicas de cada variável numérica.

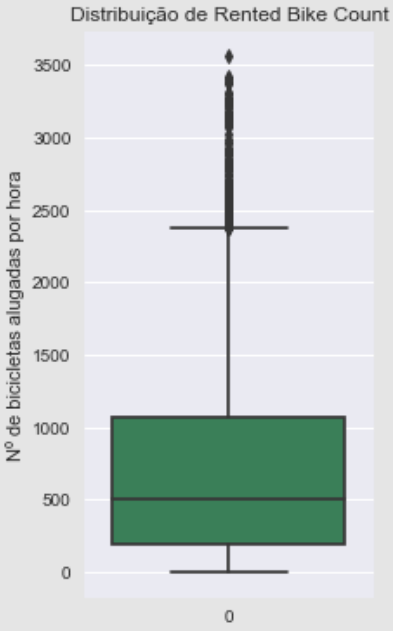
	count	mean	std	min	25%	50%	75%	max
Rented Bike Count	8760.0	704.602055	644.997468	0.0	191.00	504.50	1065.25	3556.00
Hour	8760.0	11.500000	6.922582	0.0	5.75	11.50	17.25	23.00
Temperature(°C)	8760.0	12.882922	11.944825	-17.8	3.50	13.70	22.50	39.40
Humidity(%)	8760.0	58.226256	20.362413	0.0	42.00	57.00	74.00	98.00
Wind speed (m/s)	8760.0	1.724909	1.036300	0.0	0.90	1.50	2.30	7.40
Visibility (10m)	8760.0	1436.825799	608.298712	27.0	940.00	1698.00	2000.00	2000.00
Dew point temperature(°C)	8760.0	4.073813	13.060369	-30.6	-4.70	5.10	14.80	27.20
Solar Radiation (MJ/m2)	8760.0	0.569111	0.868746	0.0	0.00	0.01	0.93	3.52
Rainfall(mm)	8760.0	0.148687	1.128193	0.0	0.00	0.00	0.00	35.00
Snowfall (cm)	8760.0	0.075068	0.436746	0.0	0.00	0.00	0.00	8.80

Verificou-se ainda que não existiam nem **duplicados** nem **valores omissos** na base de dados.
Procedeu-se então ao estudo e visualização das variáveis de forma individual.

Rented Bike Count

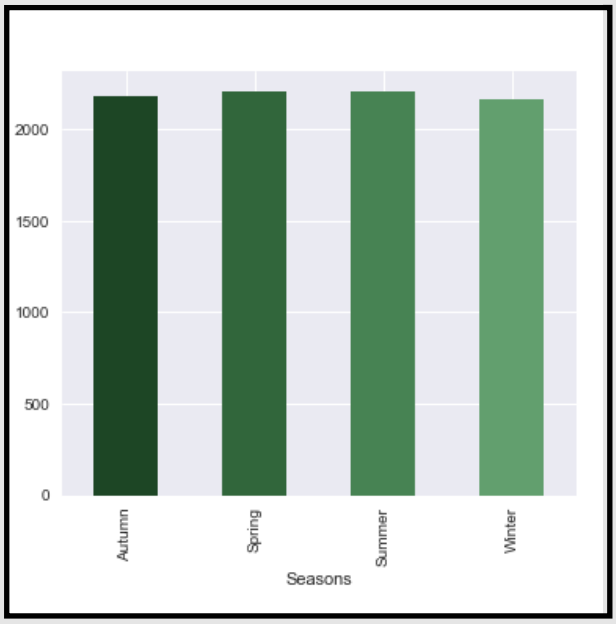
Começou-se por definir a **variável alvo ou target** do nosso estudo, Rented Bike Count que remete para o número de bicicletas alugadas no espaço de uma hora.
Seguem os gráficos que representam a distribuição da variável na base de dados:





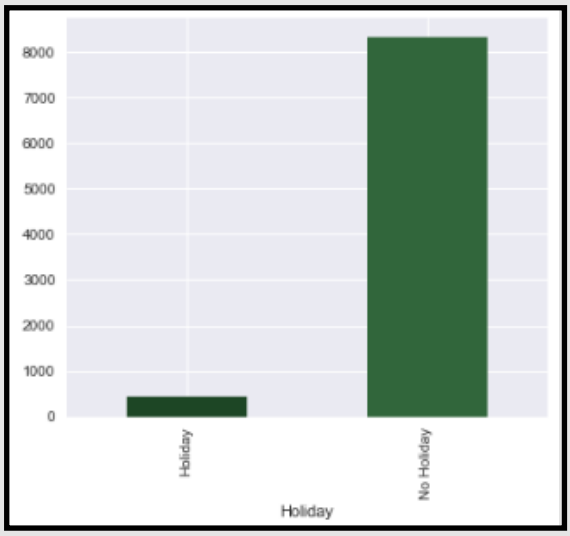
Após análise do gráfico conclui-se que na maioria das horas são alugadas entre 0 e 1200 bicicletas.
É importante notar que existem outliers superiores, nomeadamente quando são alugadas mais do que 2400 bicicletas, aproximadamente.

Seasons



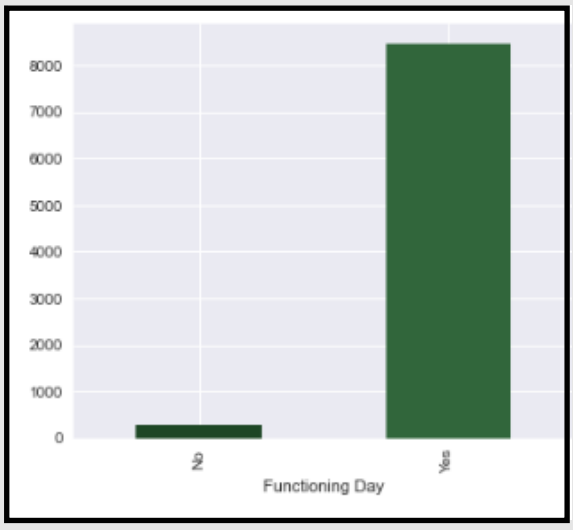
Procedeu-se para o estudo da distribuição dos dados pelas diversas classes das variáveis categóricas constituintes do dataset, começando pela variável **Seasons**.
Este permitiu verificar aquilo que já se sabia inicialmente, que as entradas da base de dados estão igualmente distribuídas pelas várias estações do ano.

Holiday



Relativamente à distribuição dos dados na variável dicotómica **Holiday** que classifica uma certa data onde foi recolhida a informação sobre o serviço de alugueres como sendo ou não um feriado nacional, conclui-se, como seria de prever, que existem mais dias em que o serviço está disponível que não são feriados nacionais do que aqueles que são.

Functioning Day

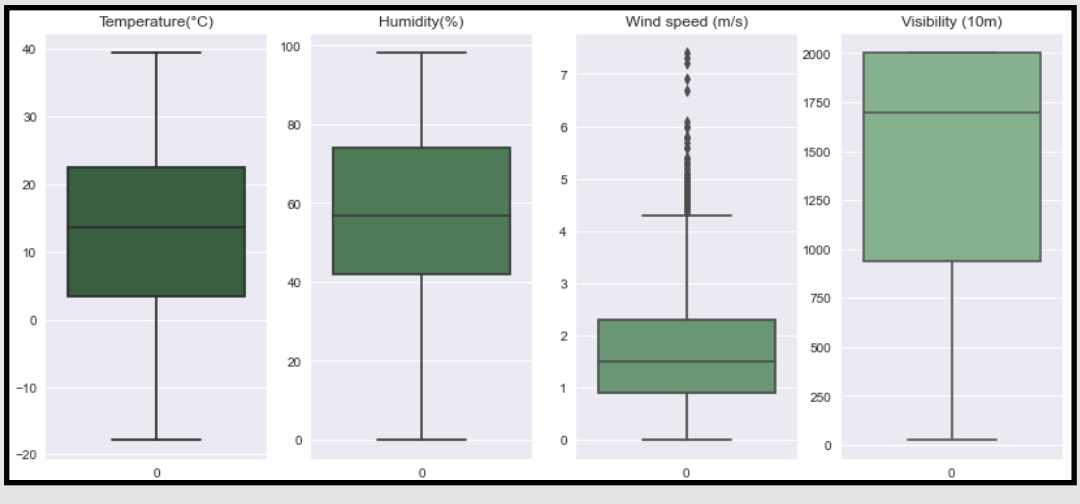


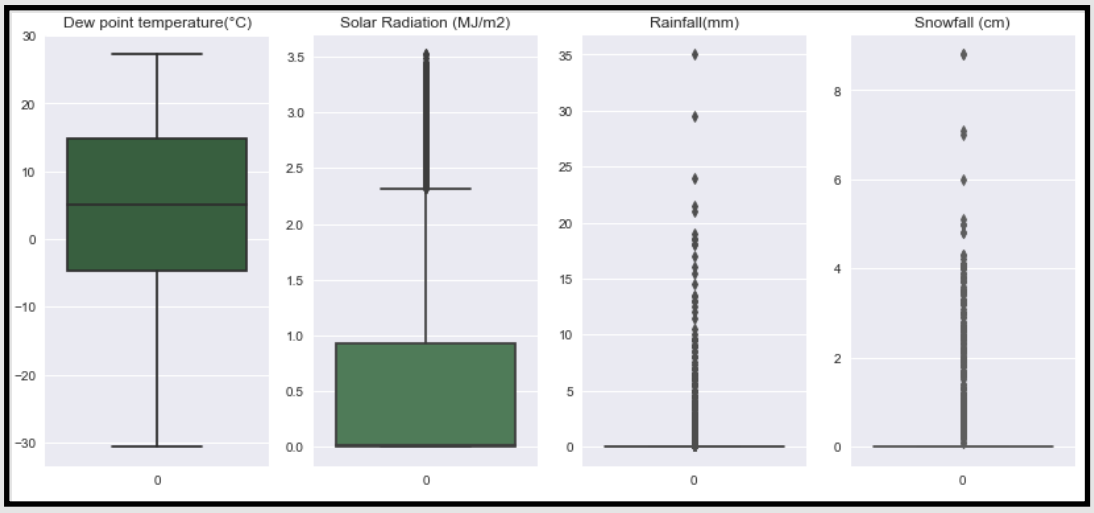
Inicialmente pensou-se que a variável **Functioning Day** correspondia à classificação das observações do dataset em dias úteis ou fins de semana, no entanto, notou-se que nos dias em que Functioning Day pertencia à categoria No, o número de alugueres por hora era nulo, chegando assim à conclusão de que esta variável classifica os dados em dias de funcionamento ou não funcionamento do serviço de alugueres.

Uma vez que esta variável não causaria impactos significantes no modelo posteriormente criado pois quando esta fosse **"No",** o número de bicicletas alugadas por hora seria automaticamente nulo e se se retirasse esta categoria **"No", Functioning Day** não teria qualquer variabilidade, optou-se pela exclusão da mesma do modelo de previsão.

Variáveis numéricas

Passou-se à visualização da distribuição de variáveis numéricas, para averiguar a existência de valores extremos que pudessem afetar a solução. Por observação gráfica, constata-se que apenas **Wind Speed** possui outliers moderados e severos superiores.





Das restantes variáveis numéricas, possuem outliers as variáveis **Solar Radiation, Rainfall(mm) e Snowfall (cm)** tendo as duas últimas um grande número de valores extremos e uma distribuição irregular a ser estudada

Outliers

- **Caso de Rainfall(mm) e Snowfall (cm)**

Em termos numéricos foram encontrados 528 outliers em **Rainfall(mm)** e 443 em **Snowfall(cm)**. Apesar de ser um número elevado, não se considerou que os mesmos devessem ser eliminados da análise uma vez que constituem informação relevante para o estudo.

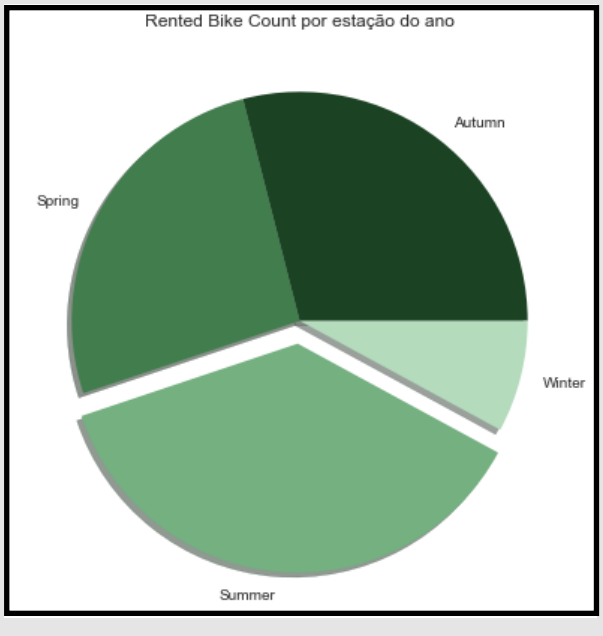
Percebeu-se que nestas variáveis todos os valores não nulos eram considerados como valores extremos e caso fossem eliminados não existiria variabilidade nem representaria a realidade correspondente às condições meteorológicas da Coreia do Sul (não chover ou nevar).

Ainda relacionado com o clima, identificou-se um padrão inesperado: na Coreia do Sul os maiores **níveis de precipitação** verificam-se no Verão, mais especificamente entre os meses de Junho e Agosto.

Seasons	Rainfall(mm)	Snowfall (cm)
Autumn	0.122756	0.056319
Spring	0.182880	0.000000
Summer	0.253487	0.000000
Winter	0.032824	0.247500

Visualização entre as variáveis

- **Rented Bike Count por Seasons**



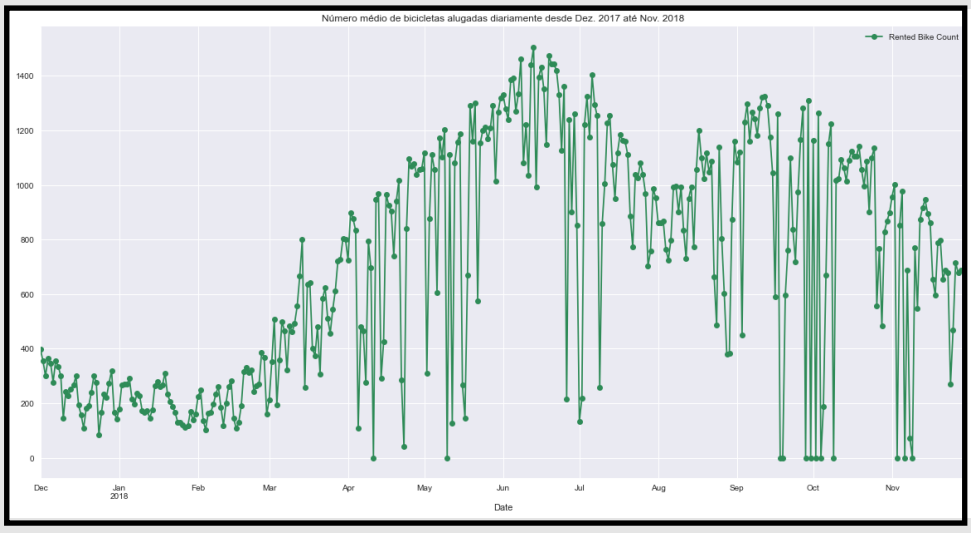
Prosseguiu-se para a visualização dos dados referentes às variáveis **Rented Bike Count** e **Season,** mais explicitamente para perceber em que estações do ano é que a procura de bicicletas era maior.

Por observação gráfica conclui-se que o maior número de alugueres ocorreu no Verão. Em sentido oposto, o menor número de alugueres ocorreu no Inverno.

- **Rented Bike Count por Holiday**

Em relação ao aluguer de bicicletas por hora consoante **o dia seja ou não feriado**, chega-se à conclusão de que o maior número de alugueres dá-se em dias que não são feriados nacionais.

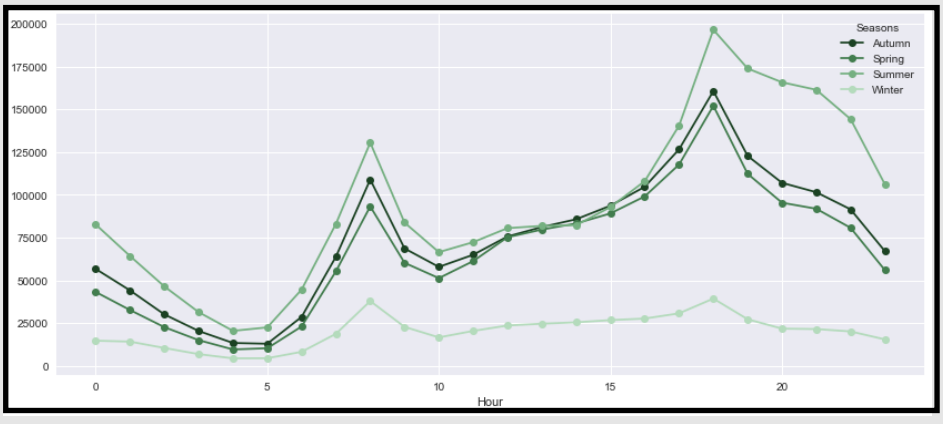
- **Rented Bike Count por mês**



De seguida criou-se uma **série temporal** que representa a variação da **média do número de alugueres de bicicletas diário** ao longo de todos os meses contidos na base de dados.

Verifica-se assim a coerência com os gráficos analisados acima referentes a que a maior procura de alugueres de bicicletas se dava no Verão, mais especificamente entre os meses de junho e agosto e vai decrescendo nos meses relativos ao Inverno.

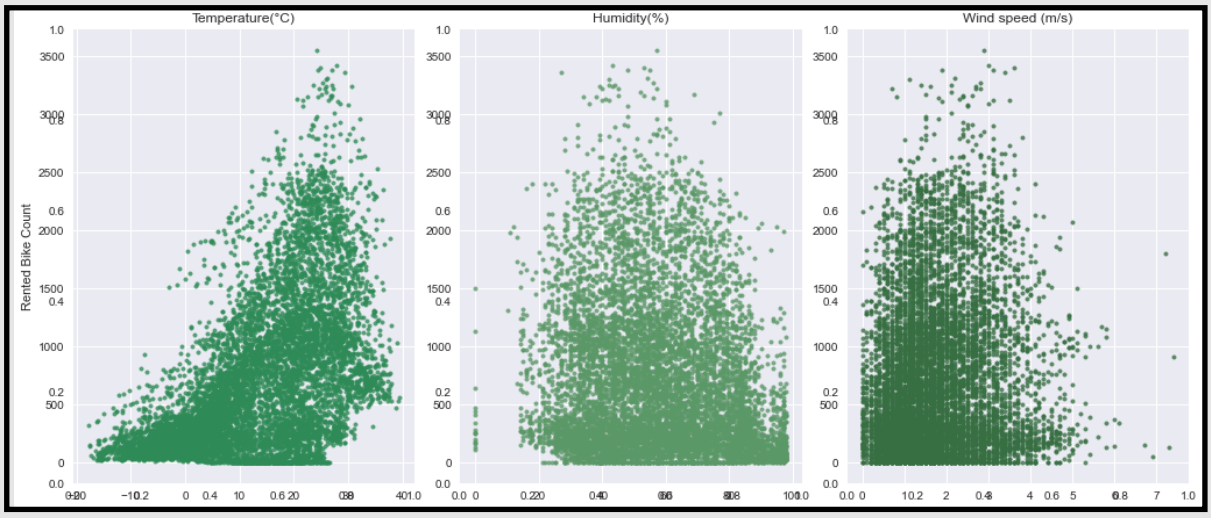
- **Rented Bike Count por Season**



Criou-se um gráfico que representa a **variação do total de alugueres de bicicletas por estação do ano** em cada hora do dia.

Nota-se que existe um padrão similar de alugueres por hora em todas as estações do ano, sendo que são atingidos picos de afluência às 8h da manhã e às 18h da tarde, correspondentes às horas de entrada e saída comuns dos trabalhadores e estudantes.

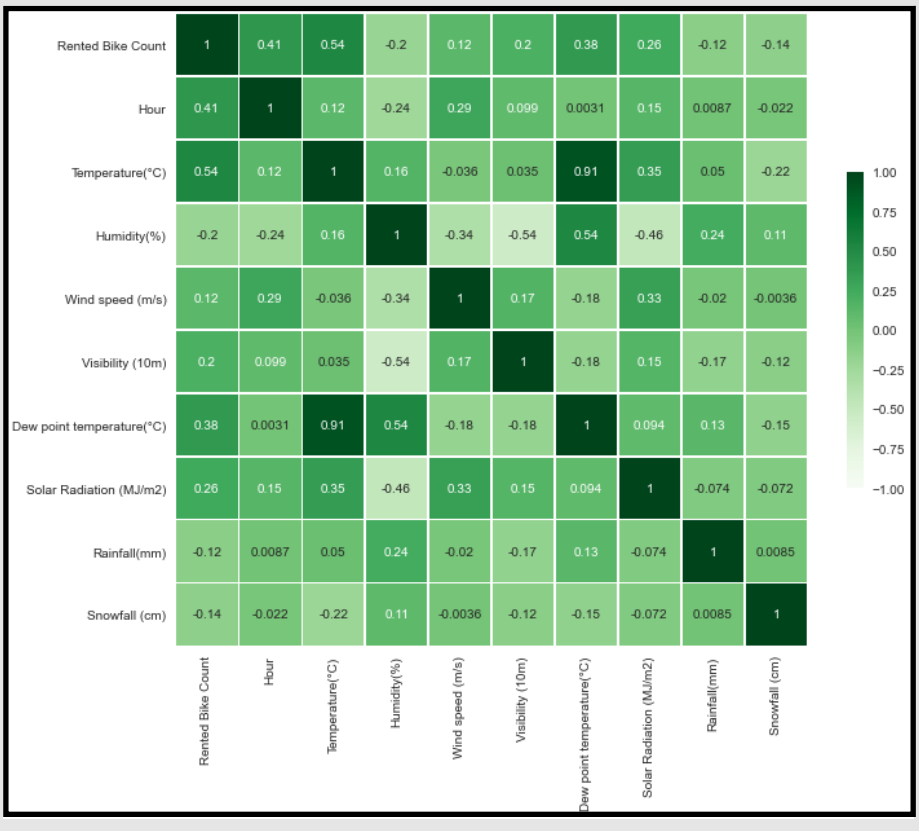
- **Rented Bike Count em relação às condições atmosféricas**



Relacionando o número de alugueres com as condições atmosféricas presentes no dataset (Temperatura, Humidade e Velocidade do vento), observa-se que:

- Em relação à **Temperatura**, esta poderá estar bastante correlacionada com a variável target na medida em que à medida que a Temperatura aumenta, também aumenta o número de alugueres por hora.
- Em relação à **Humidade** nota-se uma maior dispersão dos dados resultante numa mancha difusa de pontos, podendo significar uma baixa correlação.
- Em relação à **Velocidade do Vento** nota-se uma pequena relação entre as variáveis que parece dar-se sentido contrário, ou seja, quando a Velocidade do Vento aumenta, o nº de alugueres por hora tende a diminuir.

Correlação entre as variáveis



Seguindo para a matriz de correlações, observou-se que a variável mais correlacionada com a target é a **Temperature**, com uma correlação moderada de 0.54. A esta seguem-se as variáveis **Hour** e **Dew Point Temperature,** com as respetivas correlações 0.41 e 0.38.
Notou-se ainda a existência de multicolineariedade na explicação da variável target entre as variáveis **Temperature** e **Dew Point Temperature** através da sua elevada correlação 0.91. As variáveis **Humidity, Rainfall** e **Snowfall** estão relacionadas com a target no sentido negativo ou seja quando aumenta uma unidade seja na Humidade, na queda de chuva ou de neve, os alugueres de bicicletas tendem a reduzir.

- **Correlações das variáveis categóricas com a target**

Para além das correlações entre as variáveis numéricas do dataset e a target, efetuou-se também a análise da variância para estudar a relação entre as **variáveis categóricas** e a mesma target através da medida **Eta.**

Apresentam-se as medidas obtidas:

- **Seasons**: 3.3e-14
- **Holiday**: -1.4e-15
- **Functioning Day**: 1.113e-13

Conclui-se que nenhuma das correlações obtidas é significativa, pois estão bastante próximas de 0.
Sabe-se também que **Holiday** se correlaciona com o nº de alugueres de bicicletas por hora no sentido negativo.

Data Preparation

Prosseguindo para a fase da Data Preparation, foram tomadas ações de preparação dos dados para a posterior modelação.

- **Criação de variáveis em Escala**

Primeiramente, criou-se duas novas variáveis (**Rainfall_scale** e **Snowfall_scale**). A variável **Rainfall_scale** indica a intensidade da chuva numa determinada hora, podendo tomar as seguintes categorias:

Categorias	Intervalo (em mm)
No Rain	0mm
Light	>0 e <=2.4
Moderate	>2.4 e <= 10
Heavy	>10

A variável **Snowfall_scale** indica se estava ou não a nevar num determinada hora, sendo assim binária, com as seguintes categorias:

Categorias	Intervalo (em cm)
No	0
Yes	>0

- **Variáveis dummy**

Segundamente, construíram-se variáveis dummy para todas as variáveis categóricas presentes na base de dados (**Seasons**, **Holiday**, **Rainfall_scale** e **Snowfall_scale**) removendo a primeira dummy de cada uma.

Seasons_Spring	Seasons_Summer	Seasons_Winter	Holiday_No Holiday	Functioning Day_Yes	Rainfall_scale_Light	Rainfall_scale_Moderate	Rainfall_scale_Heavy	Snowfall_scale_Yes
8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000	8760.000000
0.252055	0.252055	0.246575	0.950685	0.966324	0.045776	0.011986	0.002511	0.050571
0.434217	0.434217	0.431042	0.216537	0.180404	0.209011	0.108830	0.050054	0.219132
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

- **Remoção de variáveis**

No que toca à eliminação de variáveis, removeram-se as variáveis **Date** (não tem relevância preditiva) e **Dew Point Temperature** (alta correlação com a variável **Temperature**, logo redundante).

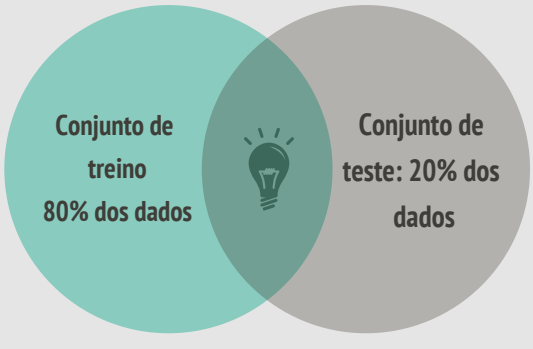
- **Estandardização de variáveis**

De seguida, procedeu-se à estandardização das variáveis com a finalidade de melhorar os resultados dos nossos modelos (no entanto apenas piorou os resultados então esta decisão foi revertida).

	Rented Bike Count	Hour	Temperature(°C)	Humidity(%)	Wind speed (m/s)	Visibility (10m)	Solar Radiation (MJ/m2)	Rainfall(mm)	Snowfall (cm)
0	254	-1.662748	-1.484762	-1.032395	0.458429	0.929577	-0.654079	-0.132495	-0.174951
1	204	-1.518249	-1.509548	-0.983575	-0.895248	0.929577	-0.654079	-0.132495	-0.174951
2	173	-1.373751	-1.550858	-0.934756	-0.701865	0.929577	-0.654079	-0.132495	-0.174951
3	107	-1.229252	-1.567382	-0.885937	-0.798556	0.929577	-0.654079	-0.132495	-0.174951
4	78	-1.084754	-1.550858	-1.081214	0.555121	0.929577	-0.654079	-0.132495	-0.174951

- **Divisão em conjunto de treino e teste**

Procedeu-se para a divisão do dataset em conjunto de treino e teste nas seguintes proporções:



Como variáveis preditoras encontram-se todas as variáveis do dataset original após exclusão das variáveis **Dew Point Temperature**, **Date** e **Functioning Day.**

Introdução à Modelação

Nesta etapa foram construídos vários modelos preditivos do número de bicicletas alugadas por hora (**Rented Bike Count**). Uma vez que se trata de uma variável alvo numérica foram utilizadas diversas técnicas de **regressão** que explicamos ao longo do relatório proposto, como a **Regressão Linear**, a **Regressão LASSO**, a **Regressão Ridge**, modelos de **Árvores de Decisão com Regressão**, **Extra Trees Regressor** e **Random Forest Regressor.** Apresenta-se a definição dos vários modelos:

- **LASSO Regression / Regressão LASSO**

O algoritmo de regressão LASSO (**Least Absolute Shrinking and Selection Operator**) é caracterizado pela regularização do algoritmo de regressão linear e seleção de preditores.
Esta regularização consiste na **redução**, a tender para zero, **dos coeficientes de regressão** mediante a adição da soma do valor absoluto da magnitude dos coeficientes (L1 regularisation) à função de perda (Soma dos Quadrados dos Resíduos) sob a forma de penalização.

$$RSS_{LASSO}(w,b) = \sum_{i=1}^{N}(y_i - (w \cdot x_i + b))^2 + \alpha \sum_{j=1}^{p} |w_j|$$

Como consequência da regularização, é expectável que um conjunto de preditores tenha os seus coeficientes reduzidos a zero, pelo que não serão adicionados ao modelo de regressão devido à sua insignificância no que toca à explicação da variável target.

É de salientar que o parâmetro α poderá ser responsável pela presença de underfitting ou overfitting. Com o aumento de α, a variância e os valores dos coeficientes sofrem uma diminuição aliviando situações de overfitting, contudo, valores excessivos de α resultarão em casos de underfitting.
Recomenda-se que α seja determinado com recurso ao método k-fold de validação cruzada, ou seja, dividir o *dataset* em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado como conjunto de teste e os k-1 restantes são utilizados como conjunto de treino. Na realização de *repeated k-fold* não é aconselhável um nº de repetições superior a 10.

- **Ridge Regression / Regressão Ridge**

Quanto ao algoritmo de regressão Ridge, este é reconhecido pela sua aplicação em cenários nos quais os preditores possuem uma elevada correlação entre si (multicolinearidade).

Semelhante ao algoritmo LASSO, realiza-se regularização sendo aplicada uma penalização à função de perda (Soma dos Quadrados dos Resíduos), embora, neste caso, se trate da adição do quadrado dos valores da magnitude dos coeficientes (L2 regularisation).

$$RSS_{RIDGE}(w, b) = \sum_{i=1}^{N} (y_i - (w \cdot x_i + b))^2 + \alpha \sum_{j=1}^{p} w_j^2$$

Resultante desta regularização, obtem-se a minimização da soma dos quadrados dos resíduos através da redução dos coeficientes da regressão, evitando que os mesmos sejam reduzidos a 0 ao contrário do algoritmo anterior pelo que não ocorre seleção de preditores.

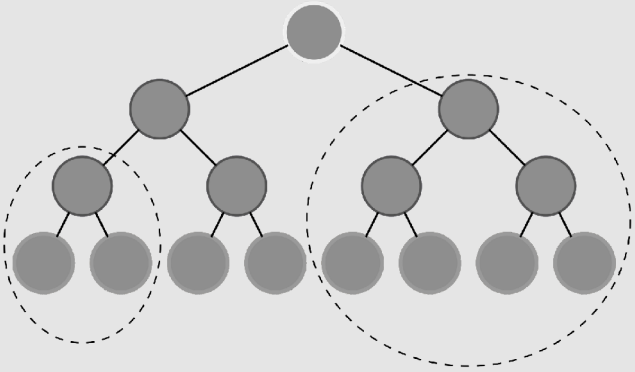
Tal como na regressão LASSO o α é arbitrário e tem um papel no alívio de situações de overfitting. Daí ser recomendável a utlização do método *leave-one-out* de validação cruzada, que consiste na criação de pares de conjuntos de treino e teste a partir do *dataset* em questão, no entanto, para este método, os conjuntos de teste de cada par correspondem a um único registo, sendo os restantes constituintes do conjunto de treino.

- **Decision Tree/ Árvores de decisão**

As árvores de decisão podem ser utilizadas para modelar problemas de classificação ou regressão. A construção destas é feita através de um modelo recursivo que divide os dados no intuito de conquistar uma boa previsão. É no nó raíz que se encontram todas as observações inicialmente.

Este nó ir-se-á dividir em 2 nós filhos ou descendentes (originando um novo nível na árvore), que posteriormente se dividirão também até alcançar os nós folha, onde serão feitas as previsões.
As árvores são modelos explicativos e preditivos que geralmente fornecem boas previsões e podem lidar tanto com variáveis qualitativas ou quantitativas. O objetivo deste modelo é sempre reduzir a diversidade da variável alvo em cada recursão.

Assim, para construirmos uma árvore de decisão precisamos de um critério que decida qual é a melhor ramificação de um nó e também de uma regra que decida quando se deve parar o processo de ramificação, resultando num nó folha ou de previsão.



O processo de ramificação das árvores normalmente escolhe, em cada recursão, a ramificação que proporciona o maior decréscimo de diversidade da variável alvo, de entre todas as possíveis.
O processo de escolha da melhor ramificação irá depender do tipo da variável explicativa naquele nó. Se a variável explicativa for numérica então teremos K-1 possíveis divisões, cada uma correspondente ao ponto médio entre os valores ordenados.
Se a variável explicativa for nominal com K categorias teremos 2^(K-1) ramificações possíveis a ser consideradas.

O processo recursivo terminará quando as regras de paragem definidas inicialmente forem atingidas. Estas regras podem incluir um número máximo de níveis na árvore ou um decréscimo de diversidade tão pouco significativo que se entende que se deve parar a modelação.

- **Extra Tree Regressor**

Ainda no domínio das árvores, foi utilizada a técnica **Extra Tree Regressor**. Esta técnica consiste na definição incial de um parâmetro *max_features* correspondente ao tamanho desejado das subamostras do dataset a utilizar.

De seguida são geradas diversas árvores de decisão nas subamostras aleatoriamente criadas de tamanho anteriormente parametrizado. O objetivo da aleatoriedade introduzida é diminuir a variância das árvores de decisão. Apesar das árvores de decisão normalmente apresentarem um melhor ajustamento ao conjunto de dados, o fator aleatório nesta nova técnica produz novas árvores de decisão com erro de previsão não tão ajustado aos dados, evitando o indesejado caso de *overfitting*.

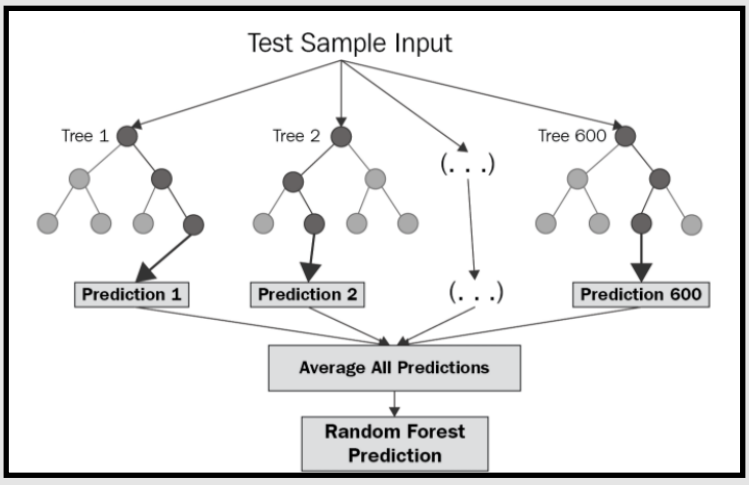
Ao fazer a média de todas as previsões obtidas, é obtido um valor mais representativo das valores preditivos. Esta técnica envolve, portanto, menos complexidade no modelo, e ao custo de um ligeiro aumento no enviesamento no modelo (menor capacidade explicativa), a variância é significativamente menor, resultando num modelo globalmente melhor, o que se veio a confirmar no dataset em estudo com a melhoria de todas as métricas preditivas (R^2, MAPE, e RMSE).

- **Random Forest Regressor**

A última técnica de construção de árvores apresentada é a **Random Forest Regressor.** É um algoritmo de Machine Learning bastante flexível e fácil de usar que produz bons resultados na maioria das vezes, gastando o mínimo de tempo no ajuste dos hiperparâmetros.

Esta técnica é responsável por aplicar um largo conjunto de **Decision Tree's**, gerando uma "floresta" em várias subamostras do dataset (de tamanho previamente definido), e utilizar a média dos valores obtidos de modo a reduzir a complexidade do modelo e evitar situações de *overfitting*.

A utilização de múltiplas árvores de decisão garante estabilidade ao algoritmo e reduz também a sua variância.



Modelling

Inicialmente, testaram-se os modelos de regressão linear simples e Regressão LASSO utilizando as variáveis não estandardizadas e estandardizadas comparando-se os resultados:

Estandardizadas

	Modelos	R^2
1	Regressão Linear standardized com escala	0.557071
2	LASSO standardized com escala	0.557022
3	Regressão Linear standardized com var. originais	0.537481
4	LASSO standardized com var. originais	0.537430

Não Estandardizadas

Modelos	R^2	MAPE	RMSE
Regressão Linear	0.557071	153.791141	437.617583
LASSO	0.557038	153.591243	437.633552

Nos modelos com as variáveis estandardizadas de início incluíram-se as variáveis em escala (**Rainfall_scale** e **Snowfall_scale**) e de seguida sem escala (**Rainfall(mm)** e **Snowfall(cm)**), percebendo-se que os melhores resultados são obtidos nos modelos que incluem estas variáveis, decidindo-se mantê-las no modelo.

Os resultados para os modelos com as variáveis estandardizadas foram de seguida comparados com os resultados dos modelos com as variáveis não estandardizadas (já incluindo as variáveis em escala). Como não foram verificados ganhos de resultados significativos a nível do R^2, decidiu-se continuar com os modelos com variáveis sujeitas a estandardização pois estes no seu indicador MAPE (erro médio absoluto em percentagem) possuem valores abolutos (número de bicicletas) sendo portanto de fácil interpretação.

Ainda assim, os resultados não foram considerados bons. Com o intuito de os melhorar, foram construídos modelos seguindo as técnicas de regressão Ridge e contrução de Árvores (Decision Tree, Extra Tree Regressor, Random Forest Regressor) sobre as variáveis não estandardizadas com escalas.

Apresentam-se de seguida os resultados obtidos:

	Modelos	R^2	MAPE	RMSE
0	Regressão Linear	0.557071	153.791141	437.617583
1	LASSO	0.557038	153.591243	437.633552
2	LASSO Afinado	0.557011	153.558042	437.646986
3	Ridge	0.557068	153.628636	437.618644
4	Decision Tree Regressor	0.770711	61.247134	314.861003
5	Extra Tree Regressor	0.884272	46.546185	223.690098
6	Random Forest Regressor	0.872616	57.412411	234.685190
7	Regressão Linear com Target Log	0.541650	73.751154	445.170030
8	Extra Tree com Target Log	0.866895	31.794533	239.896544
9	Random Forest Tree com Target Log	0.859833	35.434597	246.178798

Ainda dentro da modelação utilizando as variáveis não estandardizadas, procedeu-se à afinação dos hiperparâmetros da regressão LASSO e notou-se que não se evidenciavam melhorias significativas nos indicadores métricos.

Antes de passar à modelação por árvores, tentou-se o **modelo Ridge**, tendo-se verificado a mesma situação anteriormente referida: as melhorias foram insignificantes.

Aplicando o modelo de **árvore de decisão** aos dados**,** observou-se uma melhoria geral em todas as métricas preditivas, ou seja, um aumento significativo no **R^2** e uma descida nos erros **MAPE** e **RMSE**. Com esta nova informação obtida pela modelação através da árvore de decisão, concluiu-se que os métodos de modelação utilizando estruturas de árvore possam ser a abordagem mais adequada para estes dados. Assim foram pesquisados mais métodos relacionados com o método das arvores de decisão, tendo sido encontrado o método **Random Forest Regressor** e **Extra Tree Regressor**.
De seguida foram testados 2 modelos com estes métodos, e existiu uma melhoria notável em ambos os modelos perante a árvore de decisão, tendo especial atenção ao MAPE do modelo **Extra Tree Regressor** que melhorou imenso.
Como consta na figura 1, a variável target (**Rented Bike Count)** apresenta uma elevada assimetria positiva na sua escala original, pelo que se entendeu que a sua logaritmização resultaria numa melhoria dos resultados preditivos dos modelos.
Tal foi confirmado pelas performances dos 3 modelos relacionados com árvores de decisão com a target logaritmizada. Com especial destaque para o modelo **Extra Tree**, que foi eleito o melhor modelo com cerca de 0,866 de **R^2**, um **MAPE** de 31,79% e um **RMSE** de 239,89 .

Variável Events

Foi criada uma variável Events (de natureza binária) que indicará se numa determinada hora (registo) estava em curso um evento ou não, dando mais informação para a tarefa preditiva com o objetivo de melhorar a precisão dos resultados dos nossos modelos.

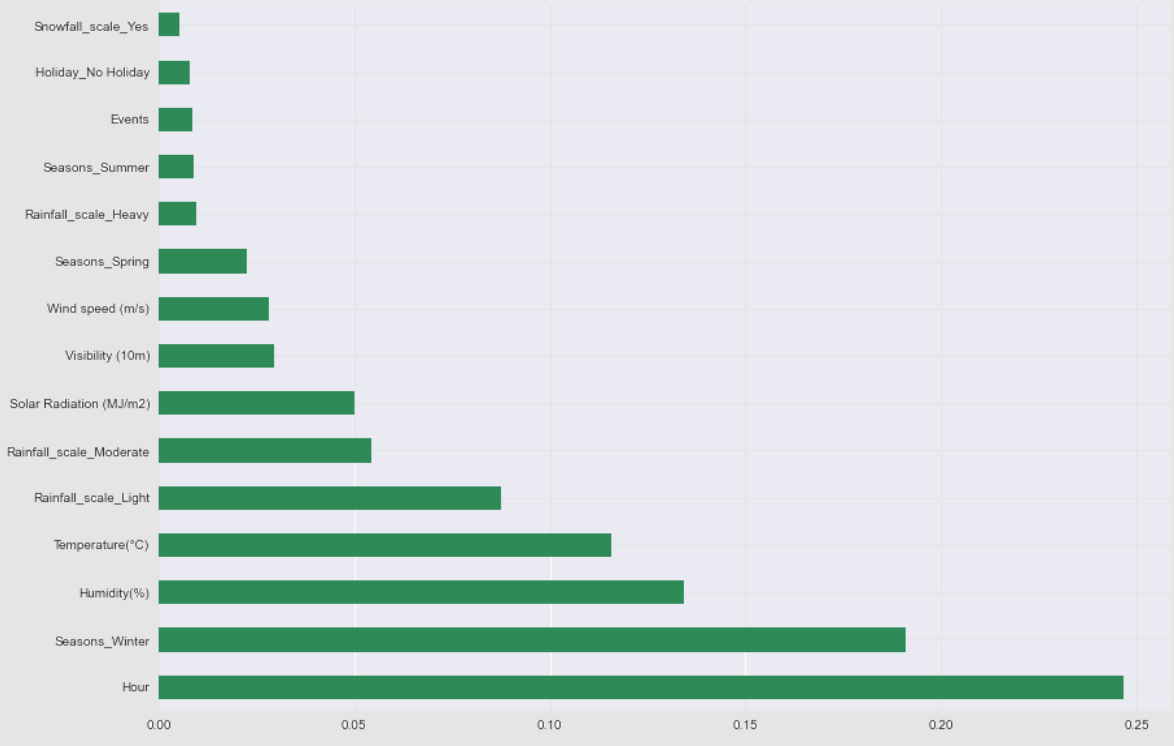
Os eventos selecionados foram os seguintes por ordem cronológica:

- Seoul Comic Word: 3 e 4 de fevereiro de 2018
- Lunar New Year Fun: 17 de fevereiro de 2018
- Protesto em Seoul: 4 de maio de 2018
- Concerto Twice: 19 e 20 de maio 2018
- UltraKorea festival: 8, 9 e 10 de junho de 2018
- Dia de eleições em Seoul: 13 de junho de 2018
- Concerto EXO: 13, 14 e 15 de julho de 2018
- Korean Liberation Day: 15 de agosto de 2018

- Concerto IKON: 18 de agosto de 2018
- Concerto BTS: 25 e 26 de agosto de 2018
- Seoul International Fireworks Festival 2018: 6 de outubro de 2018
- Concerto Monsta X: 10 de outubro 2018
- Seoul Lantern Festival 2018: 2 a 18 de novembro de 2018
- Santa Run 2018: 8 de dezembro de 2018

Notou-se que em alguns dos eventos escolhidos calhavam em dias cujo número médio de bicicletas alugadas era dos maiores, tais como o dia das eleições da Coreia ou o dia do Protesto em Seoul, pelo que se poderia esperar alguma correlação entre estes eventos e a variável alvo.

Posteriormente, realizou-se um modelo preditivo utilizando a técnica **Extra Tree com a variável target logaritmizada,** adicionando a variável Events. Constatou-se, no entanto, que esta variável tem baixa importância no modelo em questão, sendo a terceira variável menos importante. Tal poderá ser justificado pelo facto do potencial desta variável não ter sido explorado ao máximo, não tendo sido recolhidos todos os eventos ocorridos no período temporal do dataset em estudo.



Os resultados obtidos com este modelo foram ligeiramente melhores que os anteriores, apresentando um **R^2** de 87.55% e um **MAPE** de 31.15%.
Pode-se definir uma margem mínima de bicicletas como sendo a previsão feita pelo modelo adicionando ainda uma margem de erro correspondente ao erro médio absoluto MAE obtido, 137.

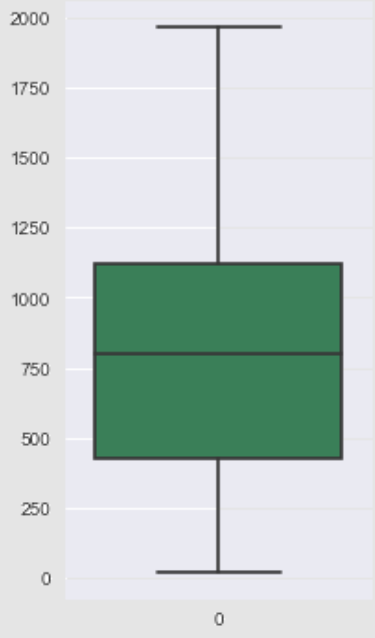
$$Nº\ Total\ de\ bicicletas\ a\ disponibilizar = previsão + 137$$

Previsão nos dias de No Functioning Day

No intuito de fornecer à empresa informação sobre possíveis ganhos em dias de não funcionamento, aplicou-se o modelo escolhido aos dias cujo "Functioning Day" era "No" e portanto, o número de bicicletas alugadas por hora era nulo.

Explorando os dias de não funcionamento da empresa, percebeu-se que maioritariamente a mesma apresentava o serviço indisponível durante as 24h, com exceção do dia 6 de outubro de 2018, em que apenas fechou por 7 horas do dia.

Os resultados obtidos a partir desta previsão variam consoante a distribuição apresentada de seguida, ou seja, entre as 400 e as 1100 bicicletas, com uma mediana de aproximadamente 800 bicicletas:



Assim, conclui-se que a empresa teria tido bastante lucro com os alugueres de bicicletas caso tivesse estado de serviço, uma vez que o número de bicicletas alugadas por hora teria sido bastante grande.

Conclusão

Chegando a esta fase do projeto, conclui-se que foram atingidos todos os objetivos iniciais propostos, ou seja, foi criado um modelo que efetua de forma precisa e coerente a previsão do número de alugueres de bicicletas numa determinada hora do dia e ainda foi definida uma margem de erro de 137 bicicletas de forma a garantir o bom funcionamento do serviço, sem que hajam falhas por escassez de bicicletas para alugar.

Dentro do modelo de previsão escolhido (Extra Tree), notou-se que a variável que mais influencia os alugueres é a Hour , que representa as horas do dia nas quais os alugueres são efetuados, sendo o intervalo temporal com maior frequência de alugueres das 07:00 ás 09:00, e das 17:00 ás 19:00. Tal deve-se ao facto destes intervalos coincidirem com as normais horas de ponta ou entrada e saída do trabalho: 08:00-09:00 e 18:00-19:00.

Também a variável Seasons_Winter (dummy da variável original Seasons onde 1 corresponde a um registo cuja época do ano é o Inverno), a segunda com mais impacto, faz notar que ser Inverno influencia bastante o número de alugueres de bicicletas a decorrer, diminuindo. Isto pode suceder tendo em conta as condições atmosféricas específicas desta estação do ano como a queda de neve, que pode impedir de forma impactante a circulação e assim influenciar bastante o serviço. É também possível demonstrar que é durante o Verão que ocorrem os maiores valores de alugueres de bicicletas.

É de salientar que inúmeros fatores podem provocar uma maior ou menor adesão ao serviço de aluguer de bicicletas, pelo que, de modo a alcançar uma boa previsão, são necessários dados que não estão presentes na base de dados original e que nem sempre se encontram ao dispor dos cientistas de dados. No caso do dataset Seoul Bikes, após uma breve contextualização quanto ao serviço que é alvo de estudo, percebe-se que não existem dados relativos à geolocalização dos pontos de aluguer e recolha, algo que permitiria o cálculo da distância (fator importante no que toca à escolha de um meio de locomoção) a percorrer entre os mesmos.

De um ponto de vista prático, suponhamos que um indivíduo se pretende deslocar da sua residência até ao local de um evento. É possível afirmar que será do seu interesse saber qual o ponto de recolha mais próximo do local do evento aquando da tomada de decisão quanto ao meio de transporte a utilizar.

Conclusão

É de extrema relevância a complementaridade existente entre os dados relativos a eventos e as geolocalizações de pontos de aluguer e recolha. Mediante a combinação das informações extraídas de ambos seria possível reforçar a quantidade de bicicletas que seria disponibilizada em pontos que se encontrem nas imediações do evento, aliviando problemas do domínio da logística.

Por outro lado no que diz respeito a eventos, seria do interesse da companhia Seoul Bikes ter conhecimento da tendência de aluguer de bicicletas não só dos dias em que estes decorrem, mas também das horas correspondentes à duração dos mesmos de modo a poder disponibilizar uma quantidade satisfatória de bicicletas aos seus clientes.

Destas necessidades resultou a criação da variável "Events", cujo potencial não está a ser explorado ao máximo dado que apenas são marcados os dias em que ocorrem eventos e não as horas em que estes sucedem. Uma análise exploratória mais profunda dentro deste domínio poderia melhorar significativamente o desempenho do modelo preditivo, pois afirma-se que o número de alugueres de bicicletas é influenciado consoante a ocorrência ou não de um evento importante em Seoul.

Por último ainda é importante concluir que os dias de No Functioning Day podem trazer ganhos significativos a nível do lucro, pelo que deve ser estudada a melhor forma de implementar os dias de não serviço em épocas do ano mais paradas, cujo número de alugueres não seja tão elevado de forma a minimizar a perda de lucro.

Bibliografia

Data Analysis of Seoul Bike Demand, *GitHub*. Acedido a 18 de abril de 2022 em https://github.com/thomastrg/SeoulBikeDemand_DataAnalysis/blob/main/final_project.ipynb

What is the CRISP-DM methodology? , *Smart Vision Europe*. Acedido a 18 de abril de 2022 em https://www.sv-europe.com/crisp-dm-methodology/#one

7 Project Success Criteria Examples (Plus Definition and Benefits), *Indeed*. Acedido a 18 de abril de 2022 em https://www.indeed.com/career-advice/career-development/project-success-criteria-examples

What Is Project Scope? 7 Steps for Defining Project Success, *Indeed*. Acedido a 18 de abril de 2022 em https://www.indeed.com/career-advice/career-development/defining-project-scope

What is CRISP DM?, *Data Science Process Alliance*. Acedido a 25 de abril de 2022 em https://www.datascience-pm.com/crisp-dm-2/

서울 (Seoul), *Ventusky*. Acedido a 26 de abril de 2022 em https://www.ventusky.com/pt/seoul

2018 Weather History in Seoul South Korea, *Weather Spark*. Acedido a 26 de abril de 2022 em https://weatherspark.com/h/y/142033/2018/Historical-Weather-during-2018-in-Seoul-South-Korea

Kumar, A. (2022). Lasso Regression Explained with Python Example, Data Analytics. Acedido a 3 de maio de 2022 em https://vitalflux.com/lasso-ridge-regression-explained-with-python-example/

Albon, C. (2017). Selecting The Best Alpha Value In Ridge Regression, Chris Albon. Acedido a 3 de Maio de 2022 em https://chrisalbon.com/code/machine_learning/linear_regression/selecting_best_alpha_value_in_ridge_regression

Singh, S. (2022). BIKE RENTAL PREDICTION, *kaggle*. Acedido a 4 de maio de 2022 em https://www.kaggle.com/code/shreya293/bike-rental-prediction

Kumar, S. (2020). Linear Regression with Logarithmic Transformation, kaggle. Acedido a 10 de maio de 2022 em https://www.kaggle.com/code/dssant85/linear-regression-with-logarithmic-transformation/notebook

Past Weather in Seoul, South Korea — Janeiro 2018, timeanddate. Acedido a 2 de maio de 2022 em https://www.timeanddate.com/weather/south-korea/seoul/historic?month=1&year=2018

How to handle time series data with ease?, pandas. Acedido a 25 de abril de 2022 em https://pandas.pydata.org/docs/getting_started/intro_tutorials/09_timeseries.html

Rain, Wikipedia. Acedido a 8 de maio de 2022 em https://en.wikipedia.org/wiki/Rain

Menestrel, T. (2022). Lasso and Ridge regression: An intuitive comparison, Towards Data Science. Acedido a 23 de maio de 2022 em https://towardsdatascience.com/lasso-and-ridge-regression-an-intuitive-comparison-3ee415487d18

Bhattacharyya, S. (2018). Ridge and Lasso Regression: L1 and L2 Regularization, Towards Data Science. Acedido a 23 de maio de 2022 em https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Bibliografia

2018 KPop Calendar, *KPOPPER'S GUIDE*. Acedido a 15 de maio de 2022 em https://kpoppersguide.wordpress.com/2019/02/02/2018-kpop-calendar/

BTS 'LOVE YOURSELF' CONCERT IN SEOUL (DAY 1), *US BTS ARMY*. Acedido a 16 de maio de 2022 em https://www.usbtsarmy.com/latest-updates/2018/8/25/bts-love-yourself-concert-in-korea

Demirkesen, Y. (2021). Applying Ridge Regression with Cross-Validation, *Towards Data Science*. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/the-power-of-ridge-regression-4281852a64d6

Wilkinson, P. (2022). An Introduction Lasso and Ridge Regression using scitkit-learn, *Towards Data Science*. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/an-introduction-lasso-and-ridge-regression-using-scitkit-learn-d3427700679c

(2011). Cross-Validation. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_190

(2011). Leave-One-Out Cross-Validation. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_469

Gurucharan, M.(2020). Machine Learning Basics: Decision Tree Regression, Towards Data Science. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

Yiu, T. (2019). Understanding Random Forest, Towards Data Science. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Tunnicliffe, D. (2021). Extra Trees, please, Towards Data Science. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/extra-trees-please-cec916e24827

Ceballos, F. (2019). An Intuitive Explanation of Random Forest and Extra Trees Classifiers, Towards Data Science. Acedido a 28 de maio de 2022 em https://towardsdatascience.com/an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b

Decision Tree - Regression. Acedido a 28 de maio de 2022 em https://www.saedsayad.com/decision_tree_reg.htm

How to prepare interactions of categorical variables in scikit-learn?, Cross Validated. Acedido a 11 de maio de 2022 em https://stats.stackexchange.com/questions/105543/how-to-prepare-interactions-of-categorical-variables-in-scikit-learn

Ensemble methods, scikitlearn. Acedido a 9 de maio de 2022 em https://scikit-learn.org/stable/modules/ensemble.html

MASE, epftoolbox. Acedido a 8 de maio de 2022 em https://epftoolbox.readthedocs.io/en/latest/modules/metrics/mase.html

Koehrsen, W. (2018). Hyperparameter Tuning the Random Forest in Python, Towards Data Science. Acedido a 9 de maio de 2022 em https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Decision Tree Regression With Hyper Parameter Tuning, NBSHARE. Acedido a 5 de maio de 2022 em https://www.nbshare.io/notebook/312837011/Decision-Tree-Regression-With-Hyper-Parameter-Tuning-In-Python/

Bibliografia

Peddisetty, T. (2020). Baby Steps Towards Data Science: Decision Tree Regression in Python, Towards Data Science. Acedido a 3 de maio de 2022 em https://towardsdatascience.com/baby-steps-towards-data-science-decision-tree-regression-in-python-323beeacbb6e

Malkin, C. (2019). Ridge Regression Python Example, Towards Data Science. Acedido a 3 de maio de 2022 em https://towardsdatascience.com/ridge-regression-python-example-f015345d936b

Ultra Korea, Wikipedia. Acedido a 17 de maio de 2022 em https://en.wikipedia.org/wiki/Ultra_Korea

Festivals&Eventos, Visit Seoul.Net. Acedido a 17 de maio de 2022 em https://english.visitseoul.net/events?srchType=&srchOptnCode=&srchCtgry=&sortOrder=&srchSchdul=C&srchBgnDe=2018-01-01&srchEndDe=2018-12-31&srchWord=

List of South Korean festivals, Wikipedia. Acedido a 17 de maio de 2022 em https://en.wikipedia.org/wiki/List_of_South_Korean_festivals#cite_note-34

Seoul Lantern Festival 2018, Visit Seoul.Net. Acedido a 17 de maio de 2022 em https://english.visitseoul.net/events/Seoul-Lantern-Festival-2018_/26573

Santa Run 2018, Visit Seoul.Net. Acedido a 17 de maio de 2022 em https://english.visitseoul.net/events/Santa-Run-2018_/27374

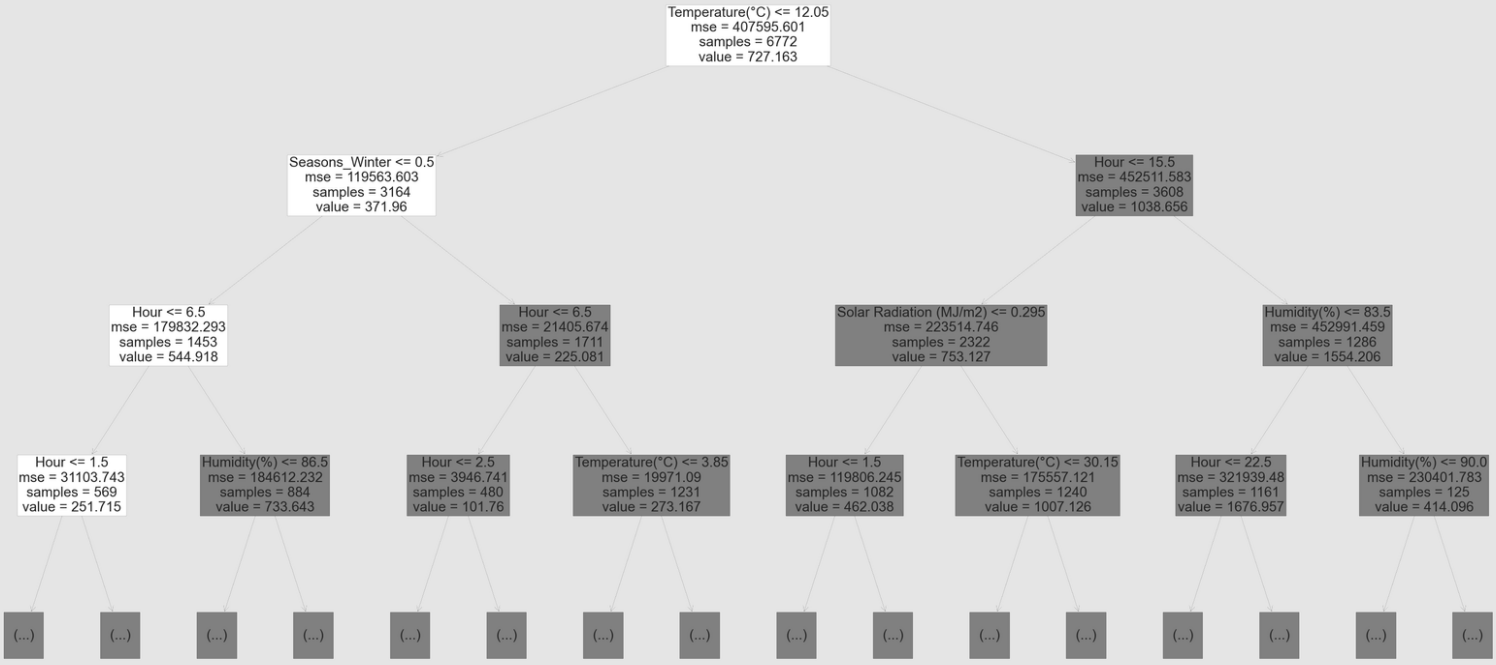
Seoul Comic World 2018, Visit Seoul.Net. Acedido a 17 de maio de 2022 em https://english.visitseoul.net/events/Seoul-Comic-World-2018_/24211

Lunar New Year Fun at the Seoul Baekje Museum 2018, Visit Seoul.Net. Acedido a 17 de maio de 2022 em https://english.visitseoul.net/events/Lunar-New-Year-Fun-at-the-Seoul-Baekje-Museum-2018_/24308

How To Change Colors For Decision Tree Plot Using Sklearn Plot_tree?, TutorialMeta. Acedido a 18 de maio de 2022 em https://tutorialmeta.com/question/how-to-change-colors-for-decision-tree-plot-using-sklearn-plot-tree

color.tree.plot: color.tree.plot, RDocumentation. Acedido a 18 de maio de 2022 em https://www.rdocumentation.org/packages/iteRates/versions/3.1/topics/color.tree.plot

Anexos



Representação gráfica da árvore de decisão utilizada no modelo Decision Tree with target log