Case GAVB: Forecasting

Felipe Glicério Gomes Marcelino

Introdução e Objetivo

- O objetivo de projeto consiste em utilizar modelos de aprendizado de máquina para predizer (Forecasting) os números de acidentes no ano de 2019

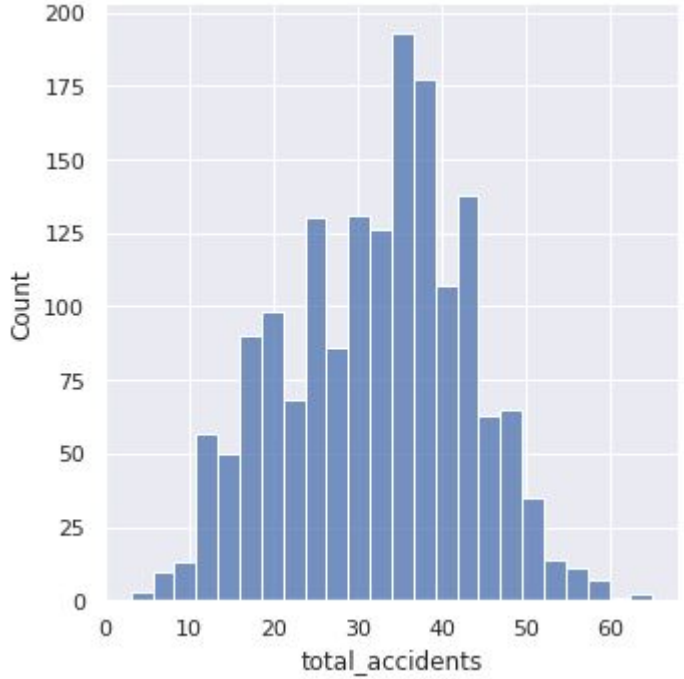
Análise Exploratória dos Dados

- Os dados é composto por um período de 2015-06 até 2019-12
  - Colunas em comum em todos os anos:
    - Data, Hora, Natureza Acidente, Situação, Bairro, Endereço, Numero, Complemento, Total de tipo de veículos presentes no acidente, Vítimas
- Na primeira modelagem utilizaremos somente **Data**
  - É o dado que está disponível em todos os anos (Veja a tabela ao lado)
  - Mas o modelo pode ser expandido (Ideias ao final do slides)

natureza_acidente	0.78
situacao	0.10
data	0.00
hora	0.10
bairro	0.54
endereco	0.41
numero	55.37
complemento	5.90
auto	7.93
moto	65.24
ciclom	87.54
ciclista	94.55
pedestre	96.59
onibus	86.57
caminhao	90.56
viatura	97.06
outros	97.86
vitimas	15.14

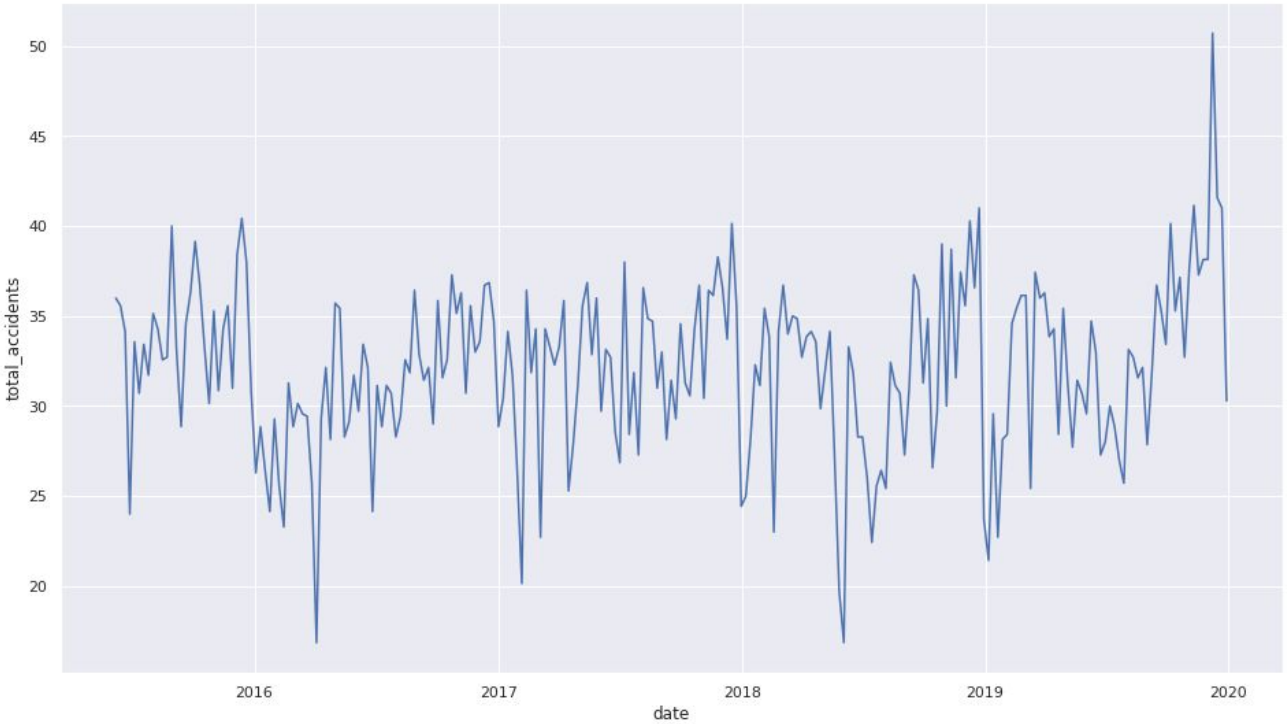
Análise Exploratória dos Dados

- Distribuição normal para os acidentes
- Não possui valores 0 e nem outliers
- Boas propiedades para aplicar MAPE



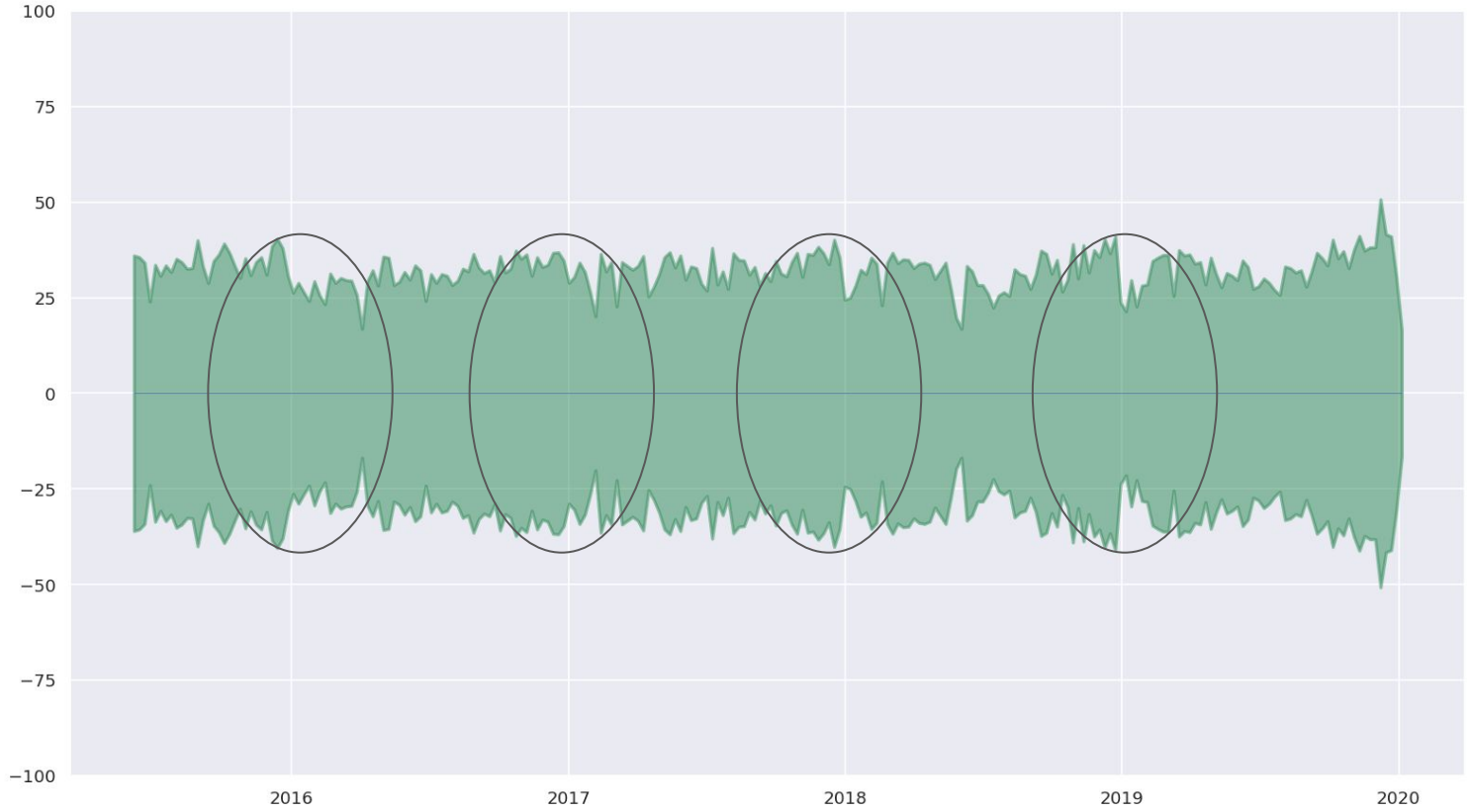
Análise Exploratória de Dados

- Média de acidentes: 32
- Mínimo: 3 (2015-06-01)
- Máximo: 65 (2019-12-31)
- Final de ano, provável que a maioria dos acidentes tenham álcool envolvido
- Será uma característica explorada pelo modelo



Análise Exploratória de Dados

Análise Exploratória de Dados

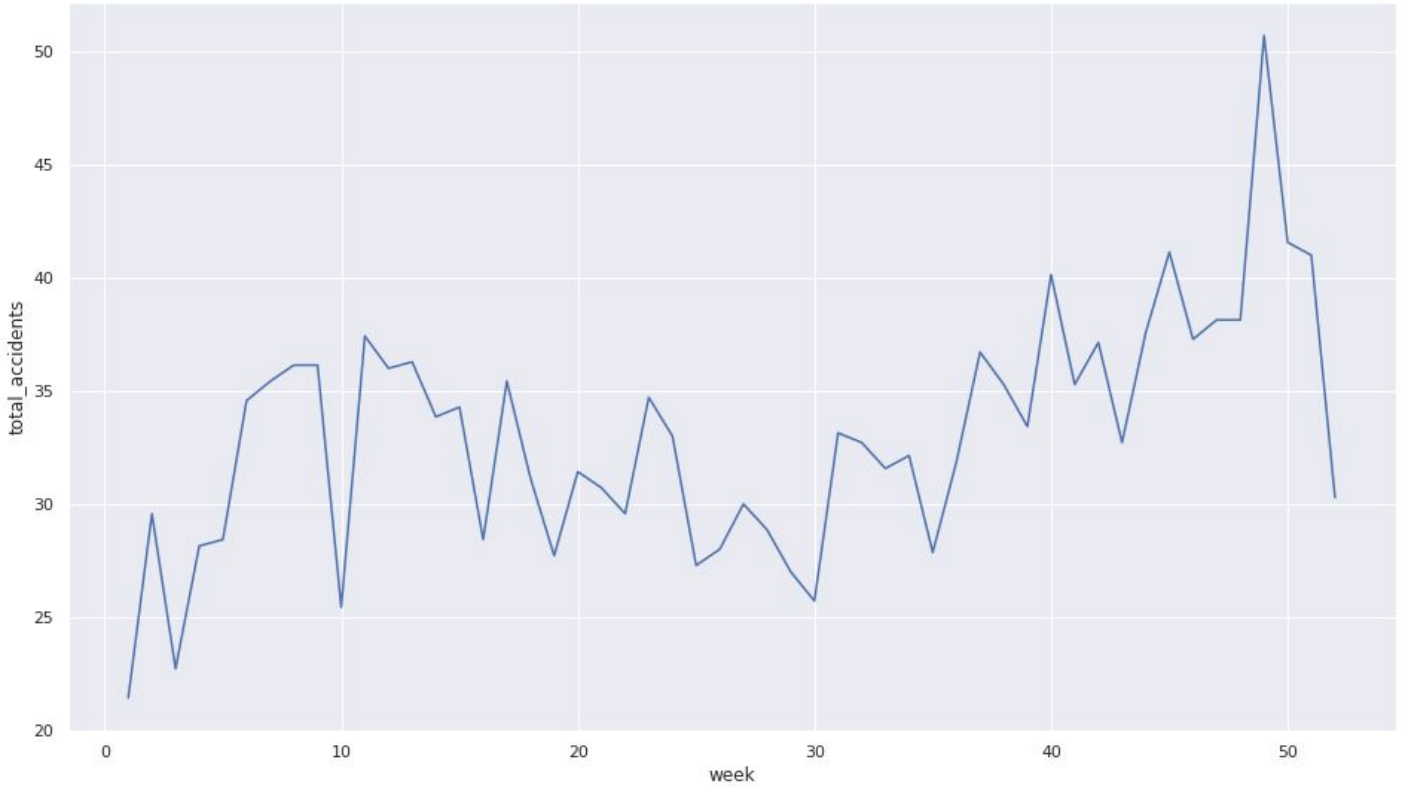


Análise Exploratória de Dados

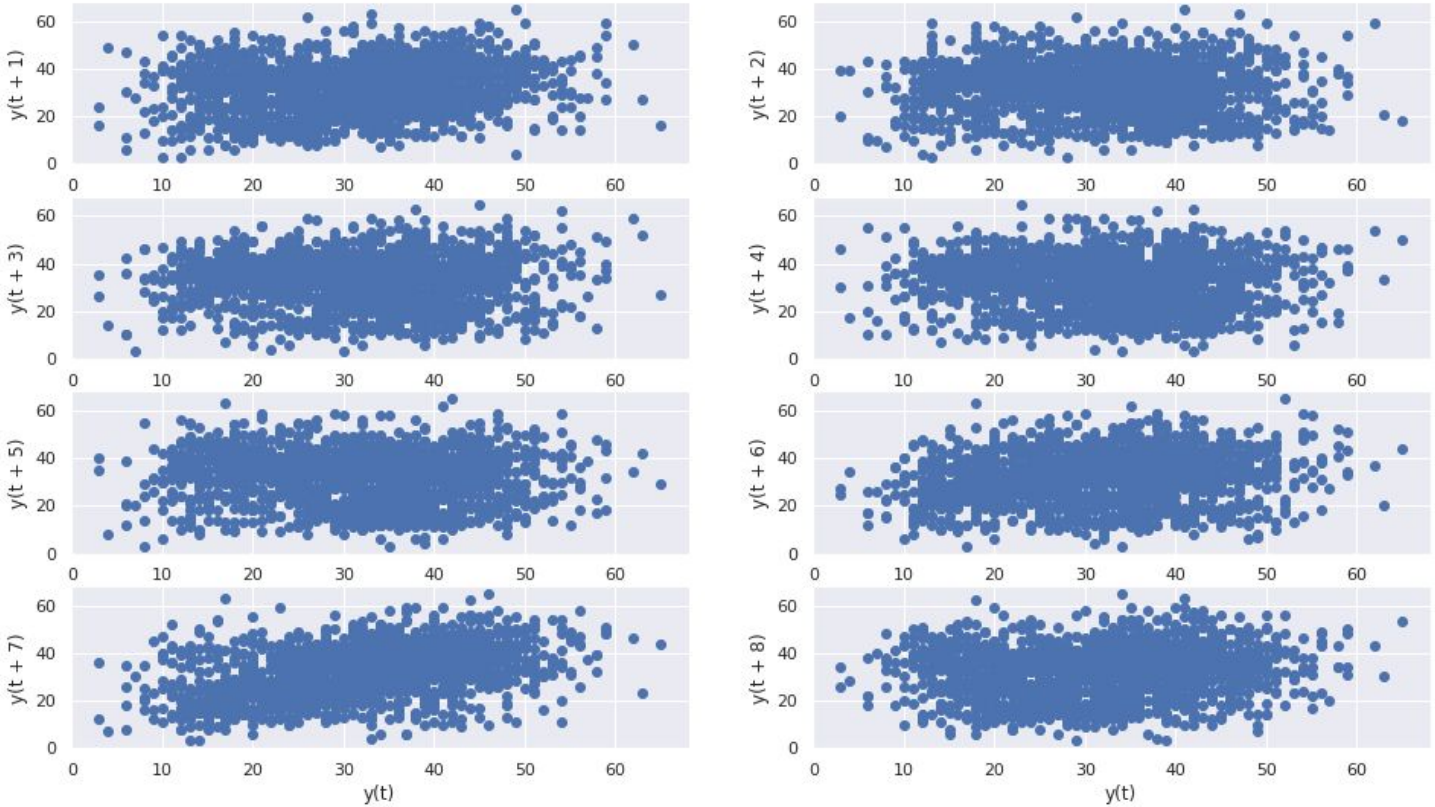
- ADF - Augmented Dick-Fuller Test

Year	ADF-p_value	ADF-Conf	Mean_p1	Mean_p2	Var_p1	Var_p2
2015-2019	6x10^10	99%	31.8	32.4	106	127
2015	0.008	99%	33.6	34.7	118	127
2016	0.011	95%	28.8	32.8	114.3	79.11
2017	0.001	99%	31.3	33.11	102.6	102.5
2018	0.014	95%	30.75	31.78	125.8	119.5
2019	**0.057**	**95%**	**31.4**	**34.7**	**125.1**	**129.3**

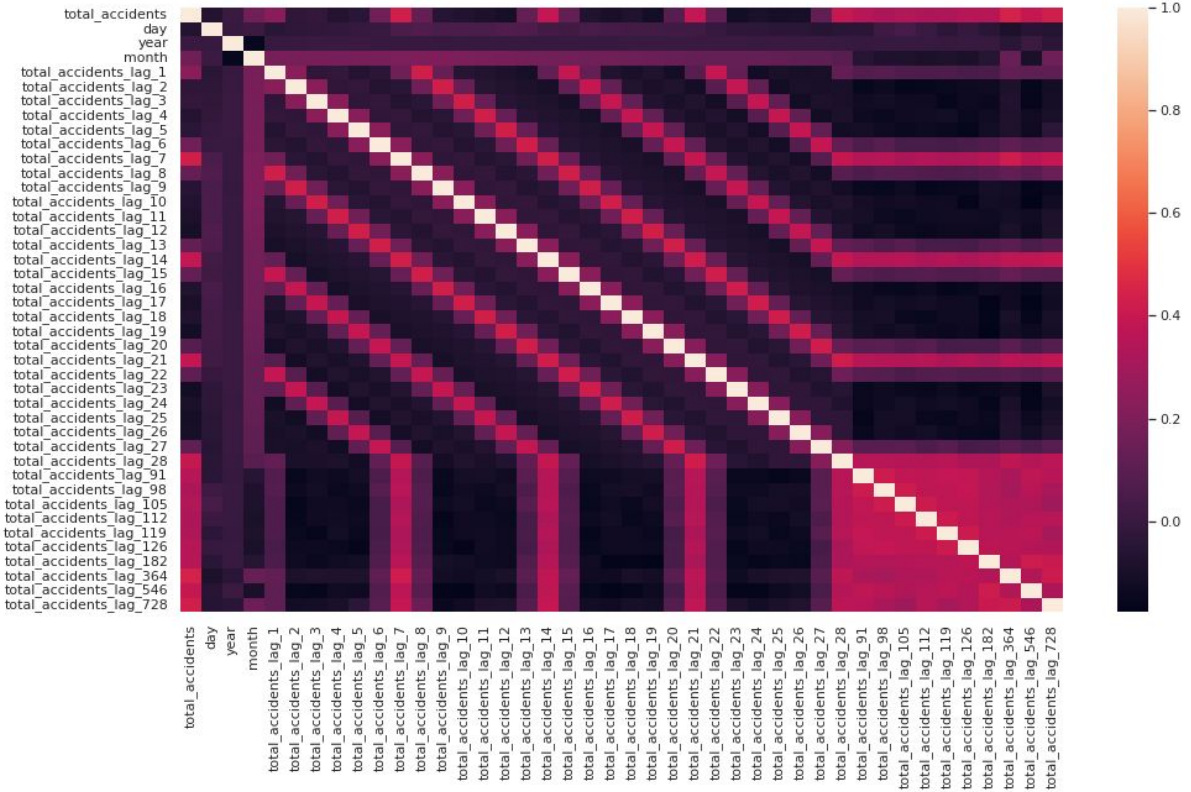
Análise Exploratória de Dados



Análise Exploratória de Dados - Lag Plots

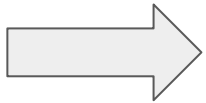


Análise Exploratória de Dados - Lag Features



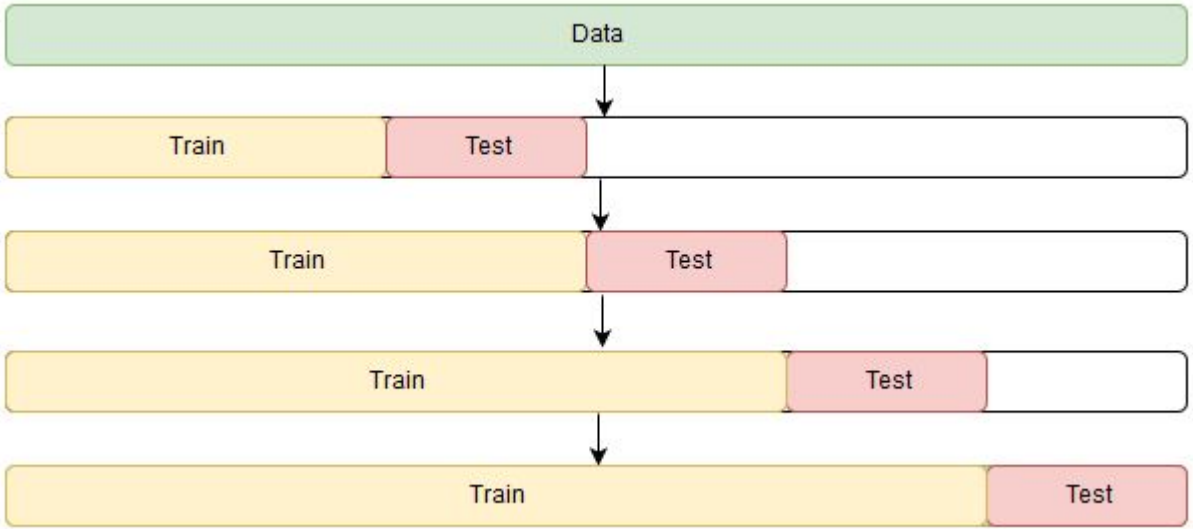
Engenharia de Features

- Week of the year
- Day of the week
- Is weekend?
- Week of the month
- Day of the year
- Dummies
  - Day of the week
  - Month
  - Week of the month
- Lag Features



Total de 1675 amostra e 69 colunas!!!

Modelo - Validação



Modelo - Seleção

- Algoritmos: XGBoost, XGBoostRF, LGBM, CatBoost, DummyRegressor
- Treino: 2015-2018
- Teste: 2019
- K-Fold: 10
- Métrica: MAPE (Mean Absolute Percentage Error)
  - Fácil de interpretar
  - Funciona bem para o nosso problema, já que este não possui valores extremos e nem valores 0
  - Outras métricas também são calculadas
    - SMAPE, MSE

Modelo - Resultados

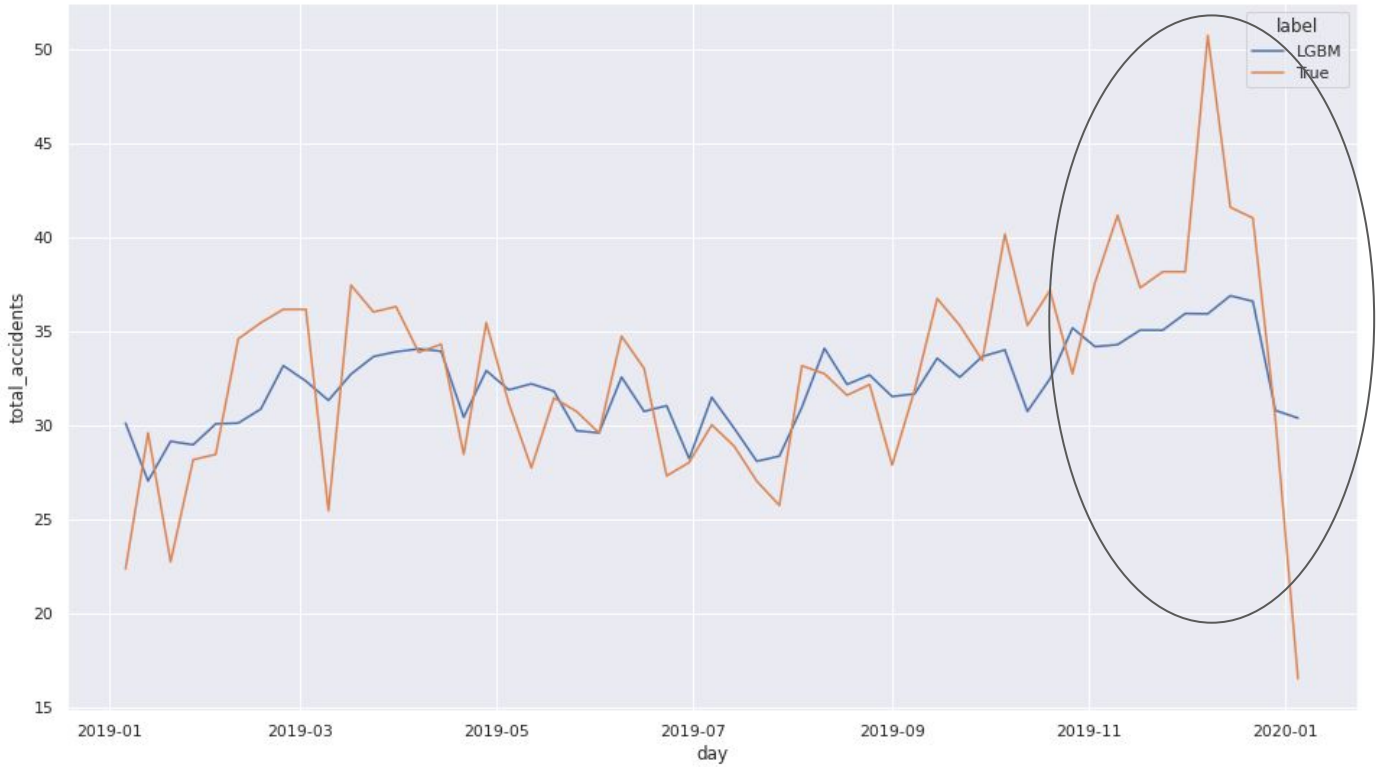
- Validação Cruzada

Algorítmo	**MAPE**
XGBoost RF	-0.2973
XGBoost	-0.3000
LGBM	**-0.2954**
CatBoost	-0.2957
DummyRegressor	-0.3955

- Métricas de conjunto de teste (Treinando nos dados 2015-2018 e testando em 2019)

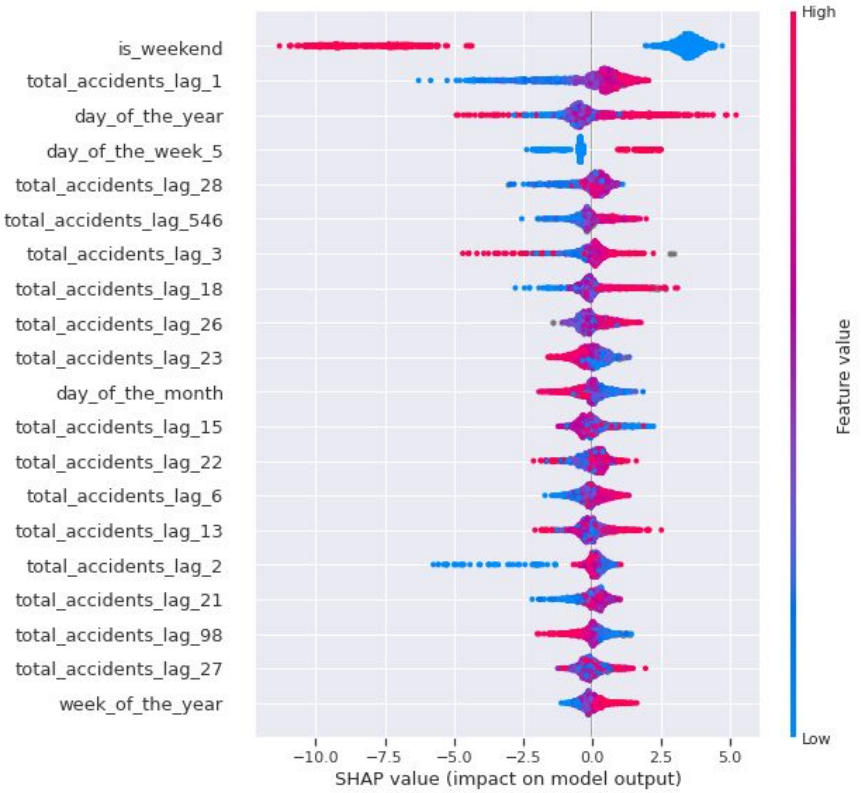
MAPE	**MSE**	**SMAPE**
75.72	0.2408	21.094

Resultados - Total por semana



Modelo - Interpretação - Shap

- Is_weekend
  - Finais de semana têm menos acidentes
- Count_lag_1
  - Quando o dia anterior possui acidentes, aumenta a possibilidade de acidente no dia atual (Provável problema estrutural ou pode ser feriados emendados)
- Day_of_the_year
  - Quanto mais ao final do ano, mais acidentes acontece. Possibilidades de mais feriados no segundo semestre (Festas)
- day_of_the_week_5 (Sábado)
  - Sábados ocorrem mais acidentes, provável que seja a mistura de festas e álcool



Modelo - Degradação

- Em torno de 10% em 5 meses
- Limite depende da aplicação de negócio e da infraestrutura/dados disponível para retreinamento do modelo

Próximos passos

- Usa transformações e diferenciações para trazer propriedades estacionária para série
- Seleção de features: backward, forward
- Criar modelos para partes específicas dos dias/regioões
  - Horário de pico
  - Maior granularidade (Regiões, bairros e etc)
- Testar modelos autorregresivos
- Imputação de dados