WOMAKERSCODE

BOOTCAMP DATA SCIENCE

2019

DESAFIO KAGGLE TRENDING YOUTUBE STATISTICS

GRUPO 05

Cristina Wada

Giseli Rodrigues

Mízia Lima

Kaggle: https://www.kaggle.com/datasnaek/youtube-new

Estudo Dataset : USvideos.csv (59.85 MB) US_category_id.json (8.3 KB)

CONTEXTO

O YouTube (o site de compartilhamento de vídeos mundialmente famoso) mantém uma lista dos principais vídeos de tendências na plataforma. De acordo com a revista Variety, “para determinar os vídeos mais populares do ano, o YouTube usa uma combinação de fatores, incluindo a medição das interações dos usuários (número de visualizações, compartilhamentos, comentários e curtidas). Observe que eles não são os vídeos mais vistos em geral no ano civil”. Os principais artistas da lista de tendências do YouTube são vídeos de música (como o famoso "Gangam Style"), performances de celebridades e / ou reality shows e vídeos virais de cara com câmera que o YouTube é bem conhecido.

Resolvemos utilizar e estudar unicamente o dataset US, como uma forma de sintetizar nosso estudo. Conseguimos observar nesse conjunto de dados as seguintes informações:

Possui **40.949 entradas** de dados, possui **16** colunas com distribuição de dados, indo desde a identificação dos vídeos carregados até informações de contagem de likes, dislikes, comentários, horários de publicação, títulos e tags, bem como os canais dos vídeos que estão subindo para a plataforma. O conjunto de dados analisado se restringiu apenas ao documento do USA (USvideos.csv), algumas perguntas surgiram inicialmente ao qual nós iremos nos ater para realizar a exploração dos dados.

Análise Exploratória dos Dados

1. Quais fatores influenciam na popularidade dos vídeos em USA?

2. Quais seriam as principais características para que um vídeo esteja entre os mais populares?

3. Consigo verificar horários e dias da semana que seriam mais interessantes pra fazer um upload com uma maior possibilidade de visualizações?

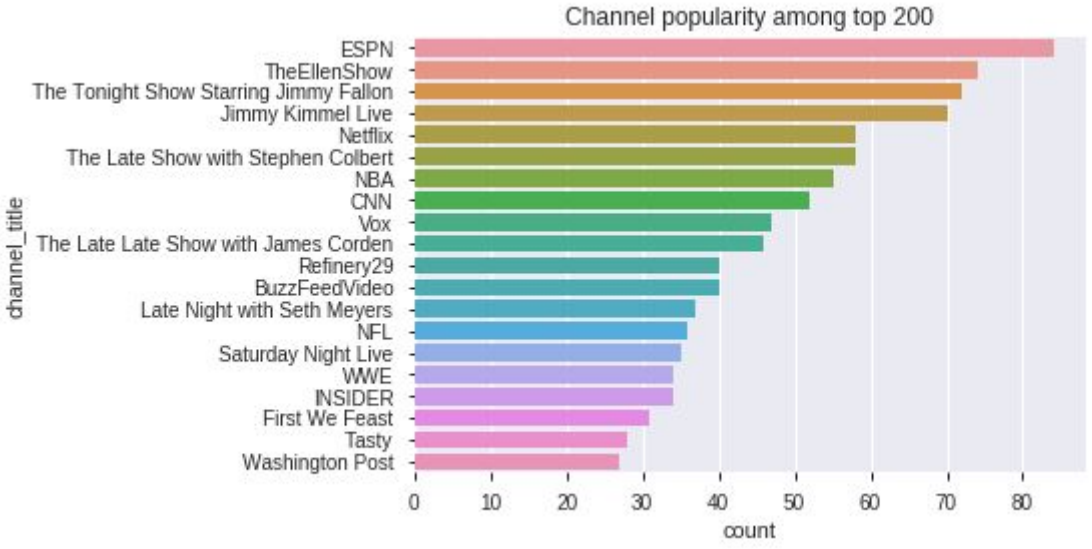
4. De que forma posso utilizar estes dados para melhorar a performance dos usuários da plataforma?

5. De que forma posso apresentar estes dados para futuros investidores da plataforma?

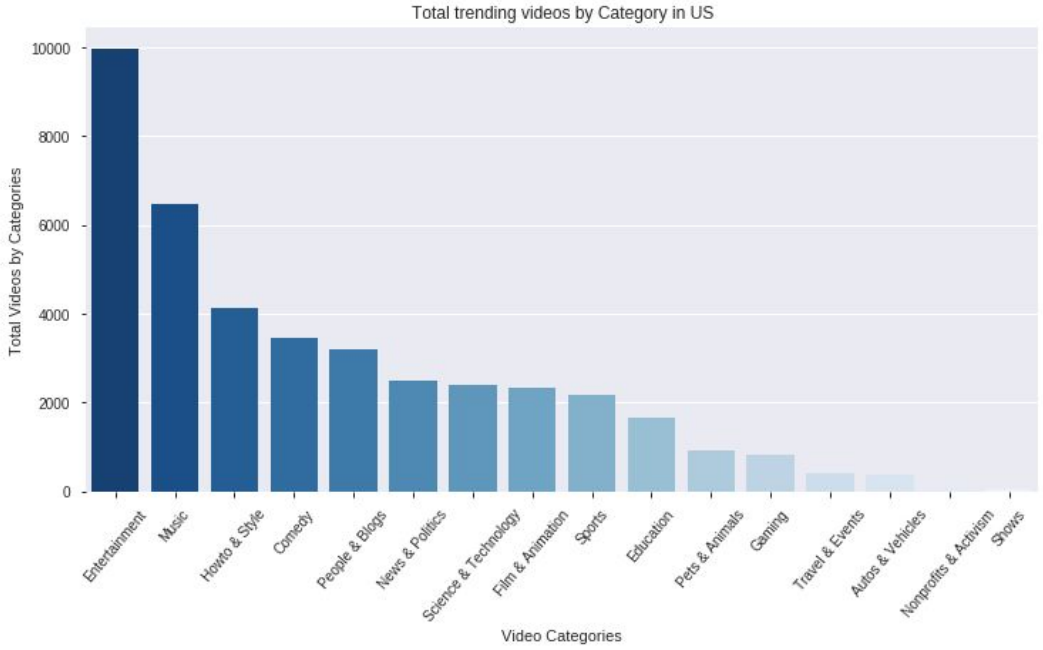
Quando apresentamos os vídeos mais visualizados no DF como um geral, aparecem vídeos nas categorias Entretenimento e Músicas, que comprovadamente mais a frente são as categorias mais populares na plataforma.

Um dos canais mais assistidos de acordo com os gráficos do DF, seria o da ESPN (programa voltado para esportes) e The Ellen Show (Entretenimento), conseguimos visualizá-los como potências de espectadores, nos levando a entender que anunciar em um

canal voltado para alguma dessas categorias ou em algum destes canais específicos, terá um maior acesso por parte dos usuários da plataforma.



Conseguimos visualizar nitidamente no gráfico de barras que Entretenimento, Música e a categoria inusitada “Howto & Style” estão entre as principais categorias acessadas, aqui conseguimos aferir que no contexto de negócios, considerando as views, likes e até mesmo dislikes (toda publicidade é “boa”), são ótimos chamarizes para anúncios, principalmente com serviços, produtos e negócios que atendam a esse público (um estudo mais aprofundado contendo média de idade, sexo, entre outros, poderia auxiliar no direcionamento de propaganda).



Observa-se também que a maior parte dos vídeos lançados na plataforma são publicados em sua grande parte nas sextas, quintas, terças, quartas e segundas tendo uma queda dessa proporção aos finais de semana. Bem como conseguimos perceber também que entre 15 e 18:00h seria o pico desses uploads, sugerindo assim esses dados que seriam ótimas informações a se passar para novos users, ou empresas que utilizam a plataforma e que querem uma maior consistência e retorno nos seus vídeos.

É possível verificar com o WordCloud, as principais palavras, tags, textos que de alguma forma estão correlacionados não só ao conteúdo mas como um chamariz para a visualização dos vídeos, uma vez que na nuvem de palavras conseguimos verificar várias destas ligadas às principais categorias listadas entre as populares, sugerindo que, se seu vídeo contém algum conteúdo que correlacione a algum destes assuntos (músicas, vídeos, trailers, assuntos oficiais, react, maquiagem, tutoriais, áudio) a probabilidade de estar entre vídeos com visualizações interessantes é maior.

Análise de Sentimentos

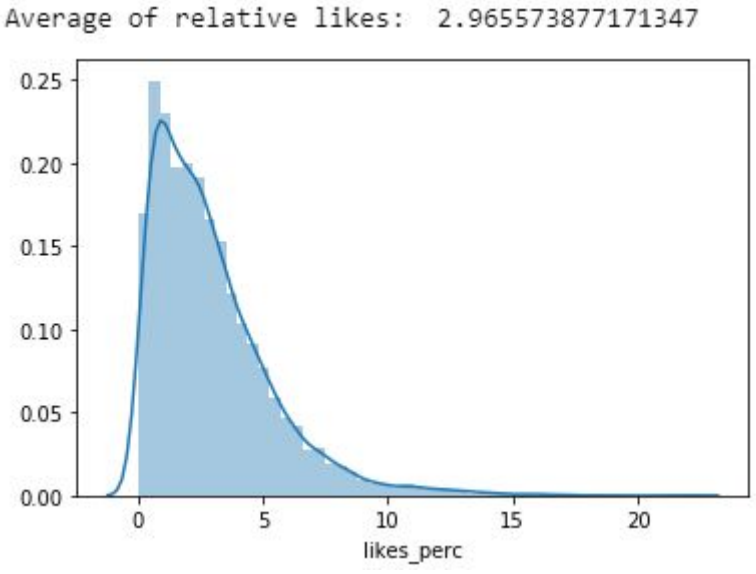
Analisar os sentimentos significa identificar se a opinião expressada em um determinado campo de estudo textual é positiva ou negativa – ou neutra. A opinião se tornou uma moeda de troca e acaba sendo um fator importante e relevante a ser considerado, já que muitas empresas se valem desse feeling com o cliente para desenvolver melhor seus produtos e serviços.

No dataset analisado, identificamos algumas features que não são relevantes pro nosso estudo, já que elas não tem interferência direta no nosso estudo. Como cada vídeo pode aparecer mais de uma vez no dataset, decidimos pegar o registro do último dia como trend e remover os demais valores ambíguos. Quantificamos então as emoções em função da proporção like, dislikes e comments_count / views.

De acordo com essas distribuições percebe-se que mesmo entre os trends vídeos, poucos usuários registram sua opinião sobre o vídeo. Iremos supor preliminarmente 4 categorias: 'loved', 'hated', 'polemic', 'neutral'

- **Loved**: vídeos com alta porcentagem relativa de curti e baixa de não curti.
- **Hated**: vídeos com baixa porcentagem relativa de curti e alta de não curti
- **Polemic**: ambas porcentagens relativas altas.
- **Neutral**: ambas porcentagens relativas baixas

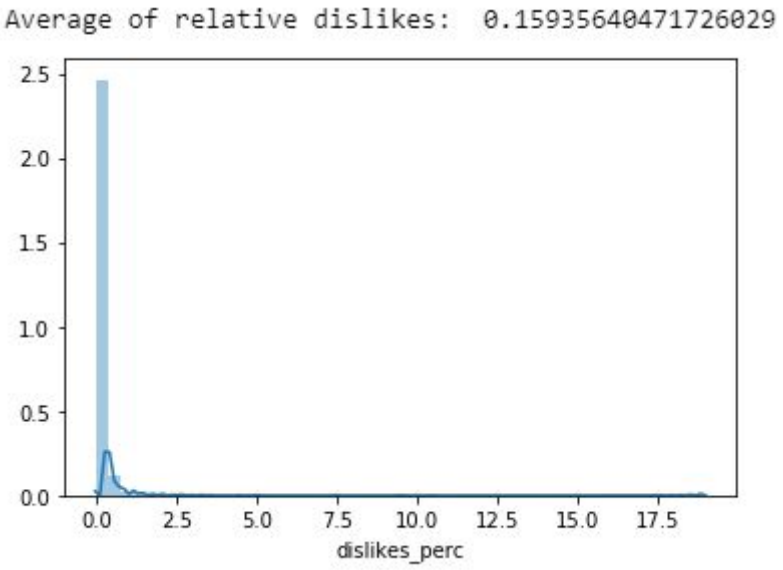
E assim conseguimos plotar em um gráfico de barras o percentual destes sentimentos, com essa solução conseguimos entender que, neste DF existe um grande percentual de sentimentos construindo para a categoria "loved" seguido por "hated"



Resultados

Realizamos uma análise sobre a distribuição de variáveis dos vídeos que potencialmente indicariam emoções por parte dos usuários. Observamos que mesmo entre os top vídeos (+ 200 milhões) de visualizações a quantidade de curtidas e não curtidas é proporcionalmente muito baixa em relação ao de visualizações. Por exemplo, em média menos de 3% das visualizações geram curtidas.

Isto indica que por si só elas não são confiáveis para apontar sentimentos, precisando de suporte de outras variáveis para caracterizar emoções que infelizmente não temos disponíveis no dataset. Esta desproporção é ainda mais forte em emoções negativas, pois em média 0.15% das visualizações de um vídeo geram não curtidas.



Tentamos criar features targets para posterior uso em algoritmos de classificação, binário (popular ou não) e multi-classe (4 tipos de emoções). Observamos pelo gráfico de barras uma desproporção na distribuição dos dados nas classes, indicando fortemente a necessidade de um pré-processamento de balanceamento. Por estas razões, os resultados dos modelos classificadores de sentimentos foram muito ruins, mesmo explorando técnicas de diversas famílias. Nas métricas os falsos positivos eram altos, indicando um enviesamento do modelo pela classe mais comum, mesmo ao aplicar balanceamento de classe.

Processamento de Linguagem Natural - NLP

O Processamento de Linguagem Natural NLP foi escolhido com o objetivo de entender e compor os textos. “Entender” um texto significa reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, e no caso iremos analisar títulos para extrair informação.

O dataset do Youtube trends possui variáveis textuais contendo informações promissoras que podem ajudar os modelos regressores ou classificadores. Entretanto, o texto precisa ser adequadamente limpo e processado. Escolhemos empregar as técnicas de NLP e regex. Expressão Regular conhecido como regex que busca identificar padrões em textos, e foi utilizado para identificar os títulos dos vídeos.

Uma vez que o texto foi processado, aplicamos o algoritmo de clusterização famoso em processamento textual, denominado Latent Dirichlet Allocation (LDA) sobre os títulos e as tags dos Youtube trend vídeos.

Feature “Title”

Investigamos a feature title do dataset. A ideia foi transformar cada título em um documento e processá-lo por NLP. Os termos tokenizados são representados em uma outra dimensão de representação. Nesta dimensão altamente esparsa, agrupamos por LDA os documentos levando em conta ao mesmo tempo:

- a frequência das palavras em cada documento
- a proximidade dos documentos de acordo com essa frequência.

Resultados

Observamos alguns padrões interessantes tanto no processamento sobre o título quanto nas tags. Os clusters confirmam o mapa de palavras mais comuns apresentado na análise exploratória e contém palavras associadas às categorias mais comuns que são Entretenimento e Música e nos canais mais visualizados.

- **Título**

cluster: ["offici", "video", "trailer", "hd"]
cluster: ["super", "bowl", "star", "audio"']
cluster: ["makeup", "test", "work", "school"'

- **Tags**

cluster: ["ellen", "movi", "trailer", "degener"]
cluster: ["makeup", "food", "tutori", "challeng"]
cluster: ["nba", "first", "game", "espn"]

Algumas regras de negócios podem ser inferidas, por exemplo, sabe-se que vídeos de maquiagem são muito populares. Entretanto o cluster indica uma frequência significativa entre os vídeos trends daqueles com termos para reviews ("test") e para maquiagem de uso no trabalho ou para estudantes.

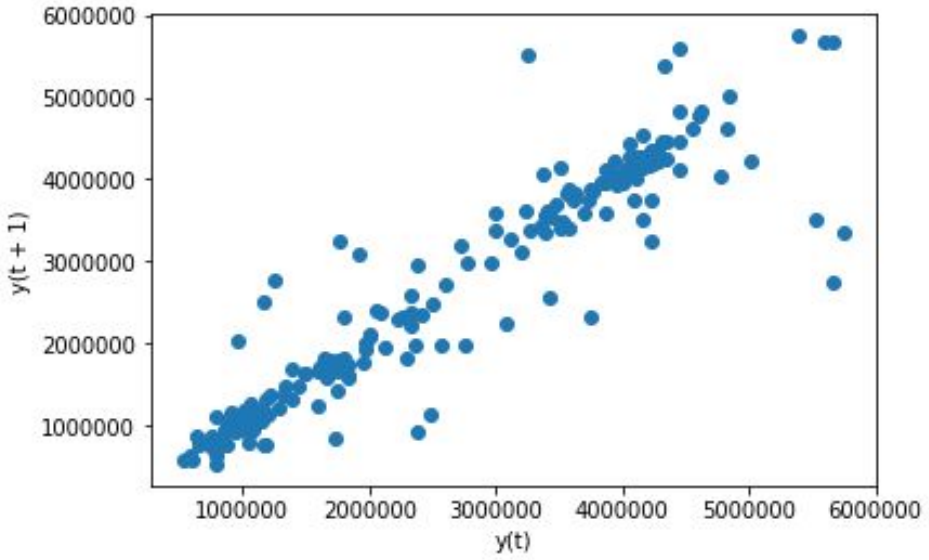
Regressão por ARIMA

Uma série temporal trata-se de uma série de observações registradas em sequência no decorrer do tempo. O dado precisa estar em frequência de tempo, ou seja, é preciso determinar o tempo dela e registrar os dados conforme essa definição (dia, ano, mês, hora).

Outro pré-requisito para uma série temporal é sua estacionariedade, ou seja, a variância e a média dela precisam estar na mesma frequência (constante) ao longo do tempo observado.

O dataset do Youtube trends possui algumas features interessantes que variam com o tempo, como a quantidade de visualizações, curtidas, não curtidas. Potencialmente, o seu comportamento pode ser capturado por um modelo regressivo e assim, podemos realizar algumas predições sobre a tendência da variável. Utilizaremos o modelo ARIMA para realizar a predição da evolução da quantidade de likes de todos os vídeos US da categoria Entretenimento nos próximos 10 dias.

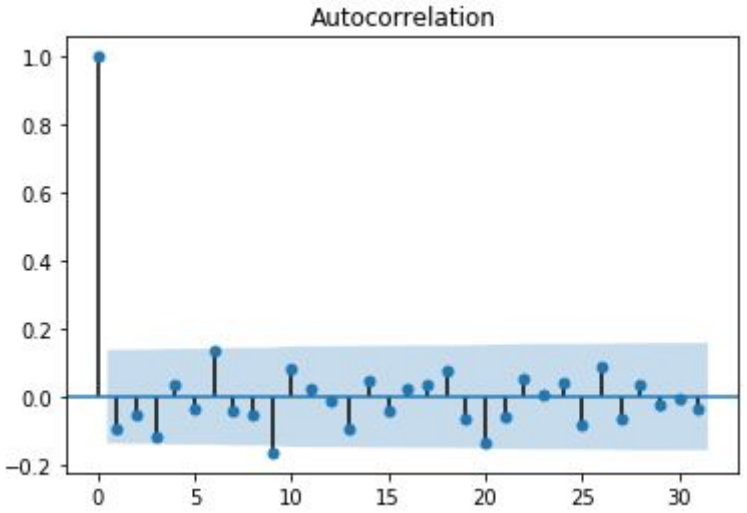
Para tal, temos o dataset US carregado, aqui vamos também utilizar o arquivo json disponibilizado e já visto na Análise e Exploração de Dados, ele possui as categorias nomeadas e que através de uma função serão vinculadas à coluna category_id. Observamos o plot destas categorias para verificar se funcionou a função, e prosseguimos para o processamento das features temporais. Utilizaremos o parâmetro datetime para tratar os dados relacionados a dias, meses e anos. Removemos as colunas não importantes para o estudo no momento e prosseguimos para verificar a evolução temporal das features avaliadas que serão: likes, dislikes, comentários e data de tendências, passando um groupby somando a função sum( ), conseguimos visualizar agrupadamente esse conjunto de informações, entretanto ainda existem algumas informações que precisam ser ajustadas no



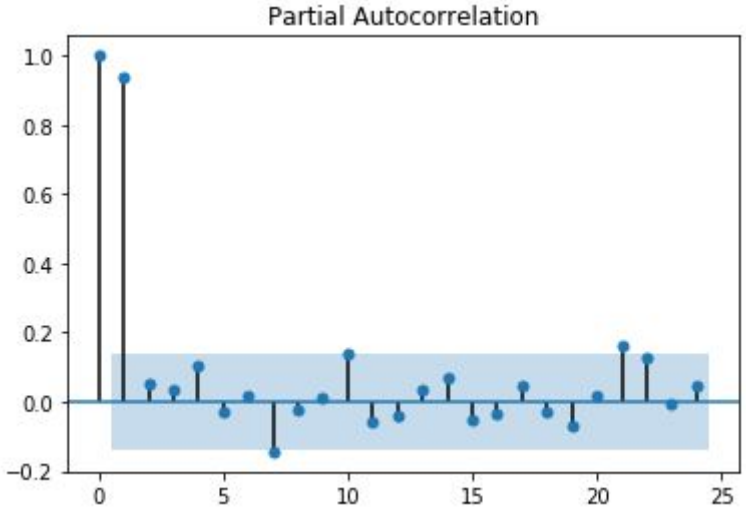
DF. Tratamos também os dados nulos da feature "likes"preenchendo com o valor "NaN" e

da feature "trending_date", com o valor da média dos valores do dia anterior, e a partir daí conseguimos verificar a evolução temporal na feature "likes".

O valor alto na diagonal secundária da matriz de autocorrelação indica uma alta correlação entre os instantes atrasados da série. Potencialmente, podemos criar um modelo ARMA para predição.

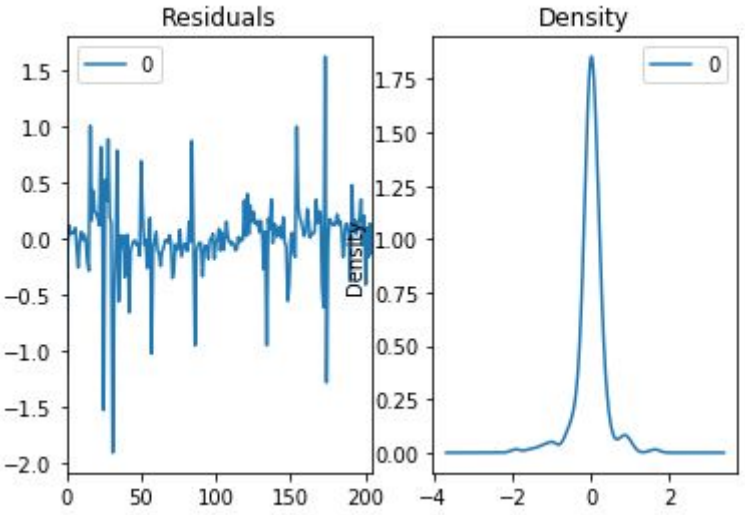


Passamos então para a verificação da autocorrelação para lags maiores, onde tempos que linha tracejada - 99%, cheia 95% são intervalos de confiança, valores de lag acima da linha apresentam significância. Uma vez apresentados nos plots, passamos a verificação de estacionariedade. Verificamos se a série está estacionária usando o teste Augmented Dickey Fuller (adfuller()). O valor p-value foi um pouco acima de 0.05 indicando que a série é quase estacionária. Aparentemente uma diferenciação é excessiva, pois lag 1 foi rapidamente para < 0. Partimos para encontrar o termo AR (p).



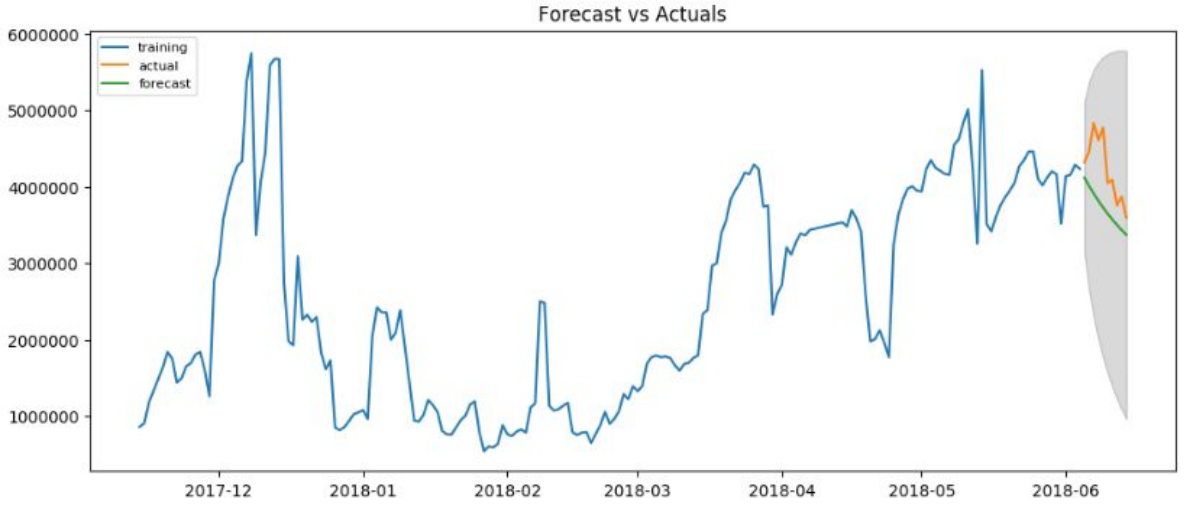
Observando o decaimento da função de autocorrelação (acf) e da parcial (pacf), temos:

- Primeira diferenciação parece excessiva, logo d=0
- 1 termo AR parece suficiente, p=1
- 1 termo MA parece suficiente, q=1



Vistos todos os termos, passamos para a construção do modelo, instanciando os valores e modelando. Outras combinações de (p, d, q) levaram a coeficientes com relevância baixa (P>|z| >> 0.05).

Apresentamos também os plots dos erros residuais, e por fim apresentamos o gráfico do modelo com a observação atual e o previsto. Foi feita também a validação cruzada out-time com a apresentação dos resultados do modelo ARMA e plot do gráfico



apresentado a curva de predição, bem como as métricas de acurácia.

Os resultados obtidos para previsão de 10 dias da evolução da quantidade de likes de todos os vídeos da categoria Entretenimento foram bem satisfatórias. O valor

predito ficou dentro da faixa de intervalo de confiança de 95% (região cinza). Algumas métricas devem ser analisadas com cautela (MAE, ME e RMSE), pois são altas mas isso não significa que o modelo é ruim, é que o range da nossa série é alta variando de (500 mil a 5.7 milhões). Por exemplo, MAPE = 0.11, indica que o modelo foi preciso em 89% na janela de 10 dias.