Data Incubator Fellowship
February 02, 2021

Presenter: Ephrem Tadesse

Title: Ethiotelecom CDR Data Analysis and Prediction

Outline

- Introduction
  - CDR dataset
  - Ethiotelecom
- Motivation
- Major features of the project
  - Visualization tools for CDR dataset
  - Exploratory analysis of mobile network download traffic
  - Prediction methods employed CDR dataset
  - Evaluation and discussion
- Significance of the project
- Target customers of the project

Introduction

- The information contained within **Call Details records (CDRs)** of mobile networks can be used to study the operational efficacy of cellular networks and behavioral pattern of mobile subscribers.
- **Ethiotelecom** is one of the government owned giant network provider company currently located in Ethiopia.
- Timesiers data analysis and tools:
  - A time-stamped dataset is sequence of data points indexed in time order
  - Various ds and ml tools such as Pandas, NumPy, Scikit Learn and clustering Models has been employed .

Introduction cont’d

- Due to high volume of demands and infrastructure limitation the government has been working to outsource Ethiotelecom for additional network providers to gain competitive advantages.
- Following this expansion, the company needs an intensive research on mobile pattern traffic analysis, spatiotemporal analysis of CDR data , temporal correlation to extract mobile traffic pattern, developing generic data-driven resource allocation approach for cellular networks based on CDR activity levels etc.

Motivation

- Motivated by this, I perform some Exploratory analysis of CDR data gained from Ethiotelecom.
- Thus, on the basis of exploratory analysis insights of some relevant features such as total call duration, call fee and network download traffic, I propose a framework for mobile network download traffic prediction corresponding to call duration , Call fee and temporal pattern .

Exploratory analysis and prediction Model selection

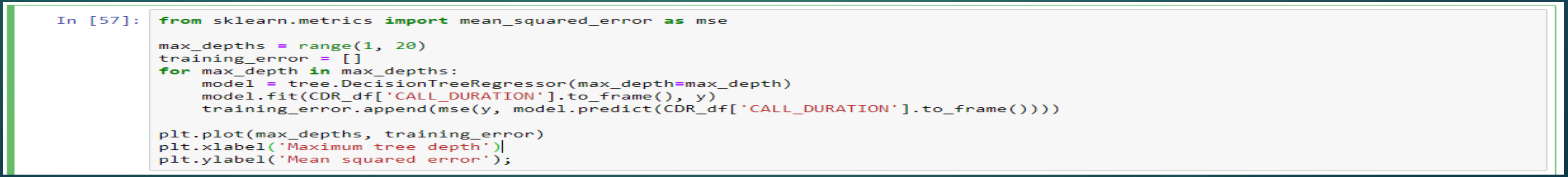
- Basic exploratory data analysis technique has been applied to extract correlation between network download patter with other feature sets.
- Most importantly important features such as call duration, call fee and temporal variations of an instance has been identified as they showed better correlation with the network download traffic.
- A decision tree algorithm has been used to predict values of network download traffic using the call duration column or feature.
- A decision tree is essentially a logic tree that branches based on feature values.

Cont’d

- By allowing for more branching, I can make the model more complex. Does this make the model better or worse?.

  (I have **changed** like from 5 to 20)

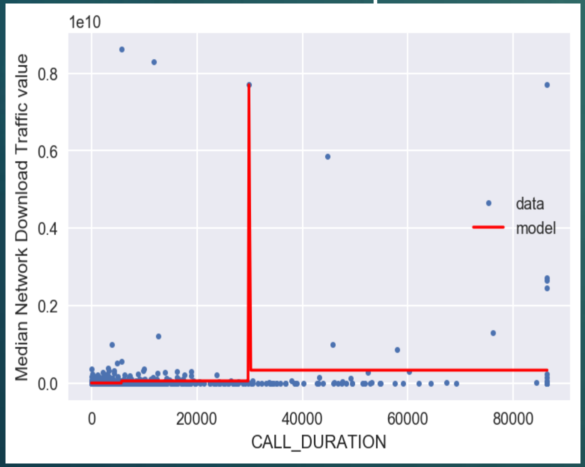
- I reach a conflict: the model looks qualitatively worse beyond max_depth > 5 but the error keeps dropping.
- This problem is called overfitting. The model looks worse because it doesn't follow the trend of the data, but instead follows the random noise.
- To detect overfitting, I need to see how our model generalizes to new data.
- I have tested this artificially by withholding part of the data set during the training step, and then using it to test the model.

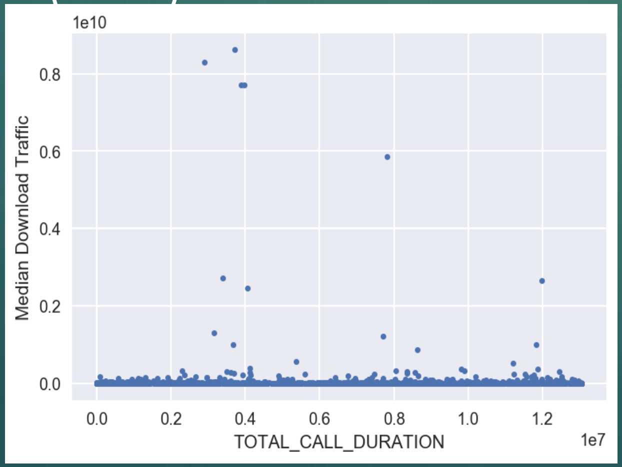


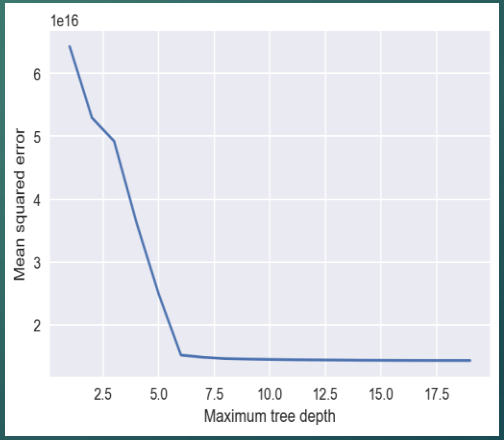
```
In [57]:  from sklearn.metrics import mean_squared_error as mse

          max_depths = range(1, 20)
          training_error = []
          for max_depth in max_depths:
              model = tree.DecisionTreeRegressor(max_depth=max_depth)
              model.fit(CDR_df['CALL_DURATION'].to_frame(), y)
              training_error.append(mse(y, model.predict(CDR_df['CALL_DURATION'].to_frame())))

          plt.plot(max_depths, training_error)
          plt.xlabel('Maximum tree depth')
          plt.ylabel('Mean squared error');
```

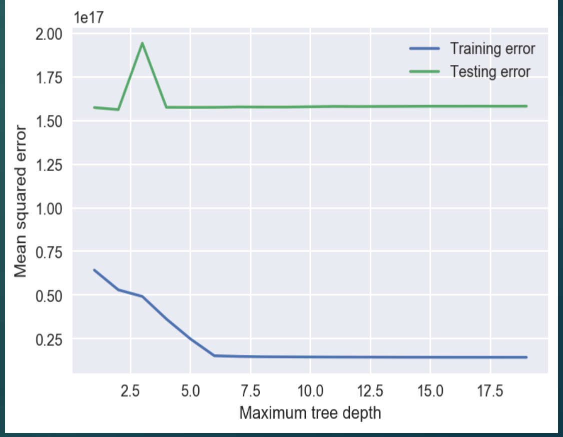
Evaluation and Discussion

- In this project, actionable insights has been extracted from the CDR data and show that there exists a strong temporal predictability in real network traffic patterns.
- Moreover the network download traffic pattern has strong correlation between the call fee, and call–duration attributes.
- This knowledge can be leveraged by the mobile operators for effective network planning such as resource management and optimization.
- Mean squared error(MSE) has been used as an evaluation metrics









Significance of the Project

- It helps Ethiotelecom to expand the network based network traffic pattern
- Competitors or other telecom company’s will invest their infrastructure based on existing network traffic sparsens
- This will help to built recommendation system for consumers and producers of telecom products and services

Target customers of the project

- Ethiotelecom
- Competitive telecom organization's who win the bids of Ethiotelecom expansion

End of presentation
Thank You!

- Presenter profile: https://ephremta.github.io/
- Project sharable link:

  https://github.com/ephremta/EthioTelecomCDRAnalysis