Plans

1. Overall list of areas/topics while building ml models
2. Pros and Cons these areas
3. Walk through of various ml case studies
4. Text mining – text classification
5. Image classification
6. Recommendation engine
7. Story telling
8. Time series (arima, prophet) – if possible
9. Stacking – if possible

Interested topics

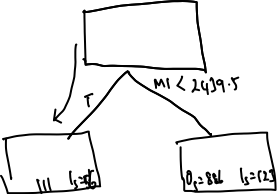
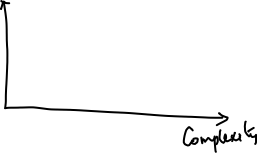
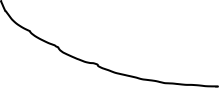
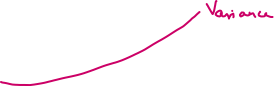
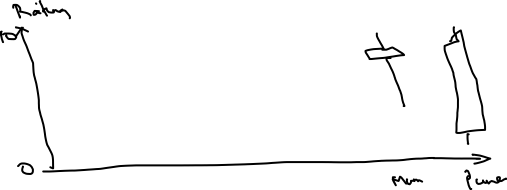
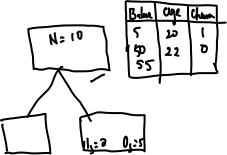
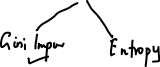
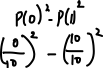
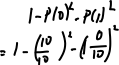
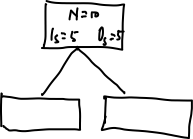
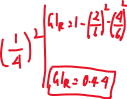
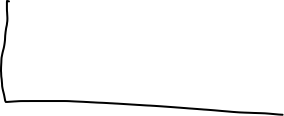
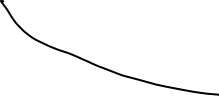
* Role of statistics in variable selection
* Missing value treatment – other methods
  + Fillna(), ffill, bfill, KNN imputer
* Bagging boosting
* Over sampling
* How tree model works for regression analysis (DecisionTreeRegressor)
* How to interpret boosting methods

Topics Covered/ Not covered

* Feature engineering (Covered) – Data type based feature engineering
* Linear regression (lasso & ridge)



* Logistic regression
* Decision Tree
  + How the tree grows?
  + Classification: Gini impurity, Regression: RMSE, MSE
  + How to plot the decision tree python? (Covered)
  + How to calculate probabilities/avg from leaf nodes in decision tree
  + Hyper parameter tuning (gridsearch, randomizedcv)
    - Training vs testing score results
    - Best\_params\_
* Random Forest
  + Variable importance
* XGBoost



**Interpretations**:

* Customers column is not there. Need to check out if we have to ignore this column. Or do some feature engg to include its properties [Feature Engg]
* Not all stores have equal no. of records. Few stores have missing sales values for certain dates [Missing records]
* When store is closed, sales is zero. Same can be applied after predicting values in the test data [Value checks, Data Manipulation]
* StateHoliday we have zero as integer as well as string. Need to combine them [Data Quality issue]. Create a new column called StateHoliday\_Flag; Value will be 1 if it is a holiday, otherwise 0 [Feature Engg]
* Create a new column, State\_School\_Holiday, by mulitiplying StateHoliday\_Flag & SchoolHoliday [Feature Engg]
* Do correlation analysis between, sales against StateHoliday\_Flag, SchoolHoliday, Promo
* Sales column is right skewed. May be apply log transformation and check if it is useful [Feature Engg]
* 2.5 years records we have in training data (2013 Jan – 2015 Jul). Predict store sales values for next 1.5 months
* December sales are higher



* Customers column is highly correlated with Sales. But we do not have Customers column in test data [Feature Engg]

**Algorithms**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Description** | **Hyper Parameters** | **Rmse** |
| Decision Tree | Base benchmark | Max\_depth = 10 | 2240 |
| Decision Tree | Customer information added | Max\_depth = 10 | 1449, 1374, 1412, 1403 |
| Decision Tree | Customer information added | Max\_depth = 12 | 1279 |
| Random Forest |  | N\_Estimators=50 | 830 |
| XGBoost |  |  | 650 |

**Variable Importance**:

A picture containing object, antenna

Description automatically generated

* C

