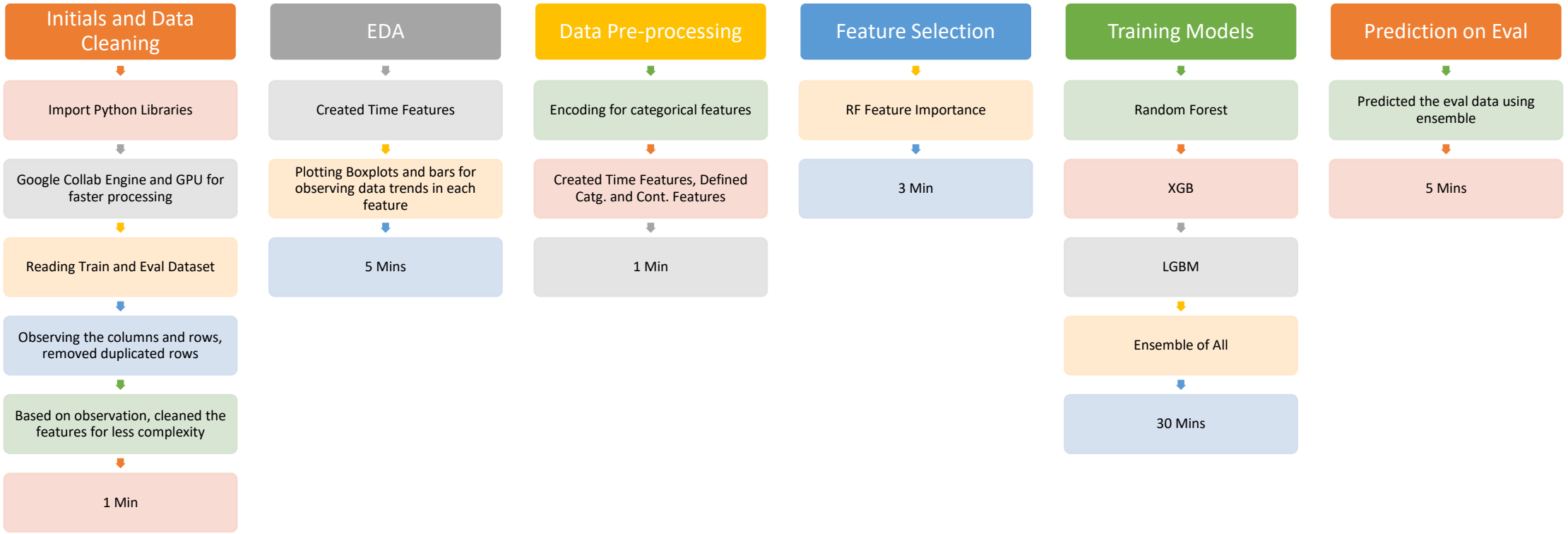
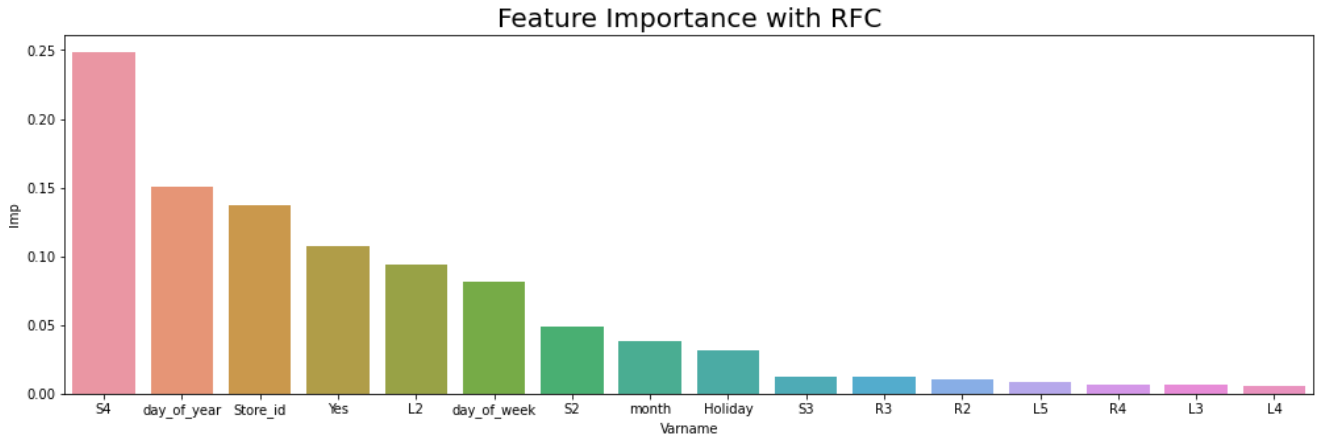
Sales Prediction Case Study : JOB-A-THON - September 2021

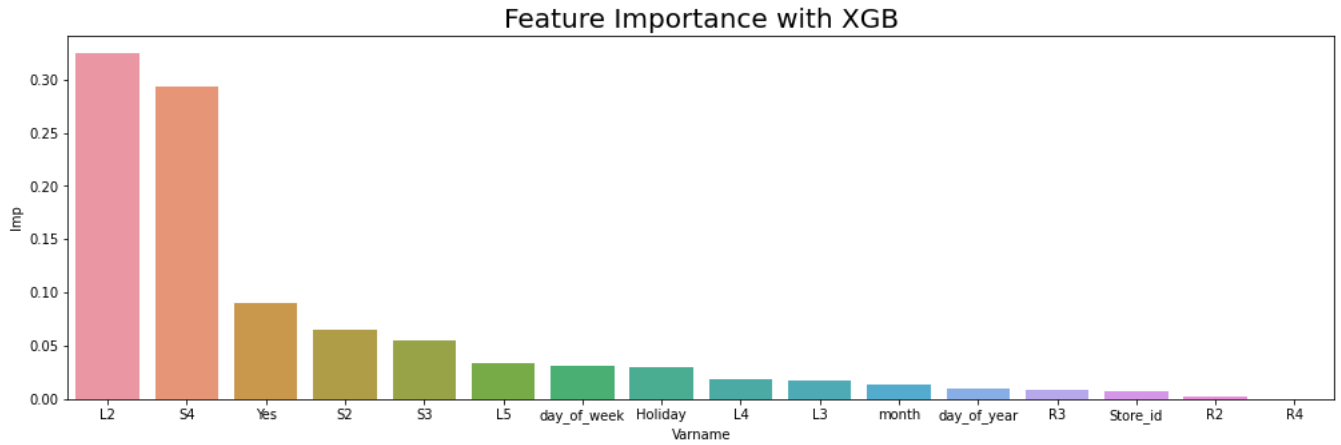
Modelling Methodology



Pre-processing & Feature Selection / Engineering

- Created Time Features using date feature
- Defined categorical variables and numerical variables
- Checked if all categories of Train and Test Datasets are matching
- Using pandas dummy feature, converted text categories to dummies features
- Changed the dtype to unit8 to make sure less memory is being used while model training
- Did not used any scaling method as tree models are being used which are independent of variance of data
- Split train dataset into training and validation datasets to check how model works for unseen data
- For important features, please refer to images attached in right.
- Union of best features from XGB and RFC can improve our score





Training Methodology

XGB

- XG-Boost is a powerful approach for building supervised regression models. The validity of this statement can be inferred by knowing about its (XG-Boost) objective function and base learners.
- **Hyperparameter:** Default
- **Training Time**: 1 Min
- **MSE**: 88.xx on Validation and 226.xx on Test

Random Forest

- Random Forests has been used for training the model, which is a collection of decision trees. The great thing about random forests is that - they almost always outperform a decision tree in terms of accuracy
- **Hyperparameter:** Default
- **Training Time**: 1 Min
- **MSE**: 66.xx on Validation and 245.xx on Test

LGBM

- LightGBM, short for Light Gradient Boosting Machine, is a free and open source distributed gradient boosting framework for machine learning originally developed by Microsoft. It is based on decision tree algorithms and used for ranking, classification and other machine learning tasks.
- **Hyperparameter:** Default
- **Training Time**: 10 Mins
- **MSE**: 64.xx on Validation and 222.xx on Test

Weighted Ensemble

- Used custom stacking for three models with different weightage of trained models to achieve lowest error score on validation and test datasets
- **Training Time**: 1 Min
- **MSE**: 64.xx on Validation and 222.xx on Test

What’s Next?

With Better Parameter tunning using Bayesian Method, Grid Search etc., optimal results can be achieved for sales prediction

New Features can be derived like which store has the highest to lowest sales in past months and so on after careful thinking and how these are affecting our final model in terms of model error rate.

Only Few features can be used to avoid overfitting after running few stats test like chi square, anova etc.