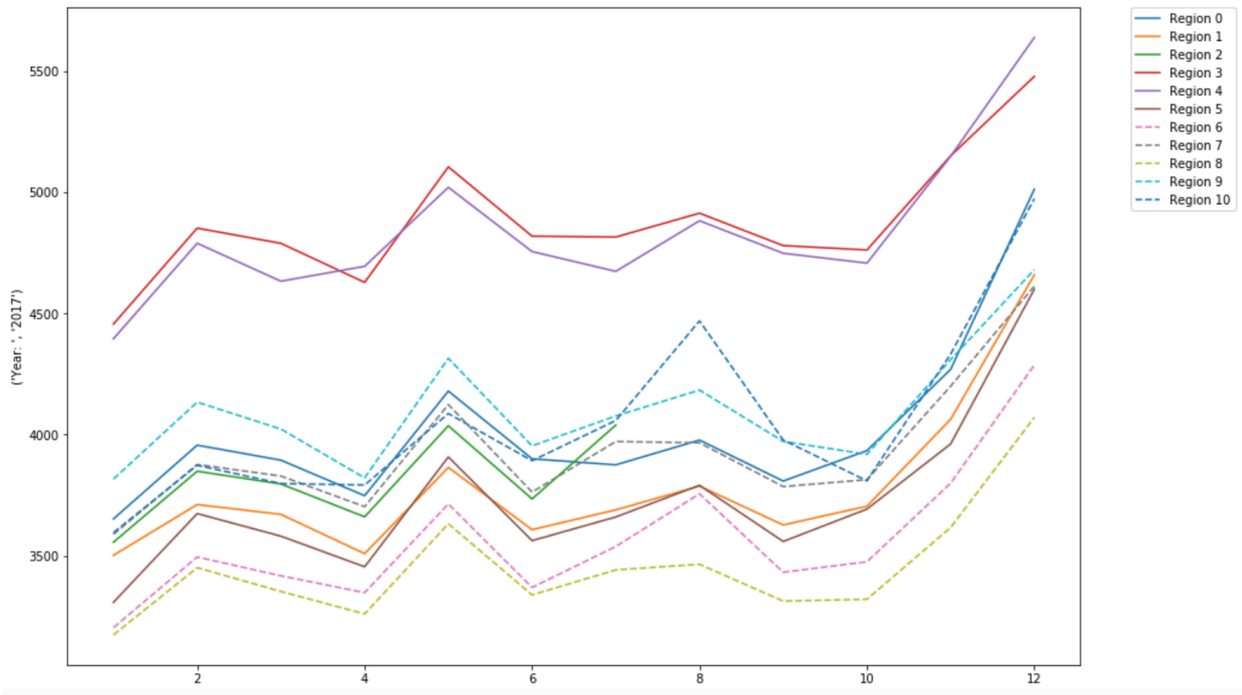
DATA MINING AND TEXT MINING

CHALLENGE

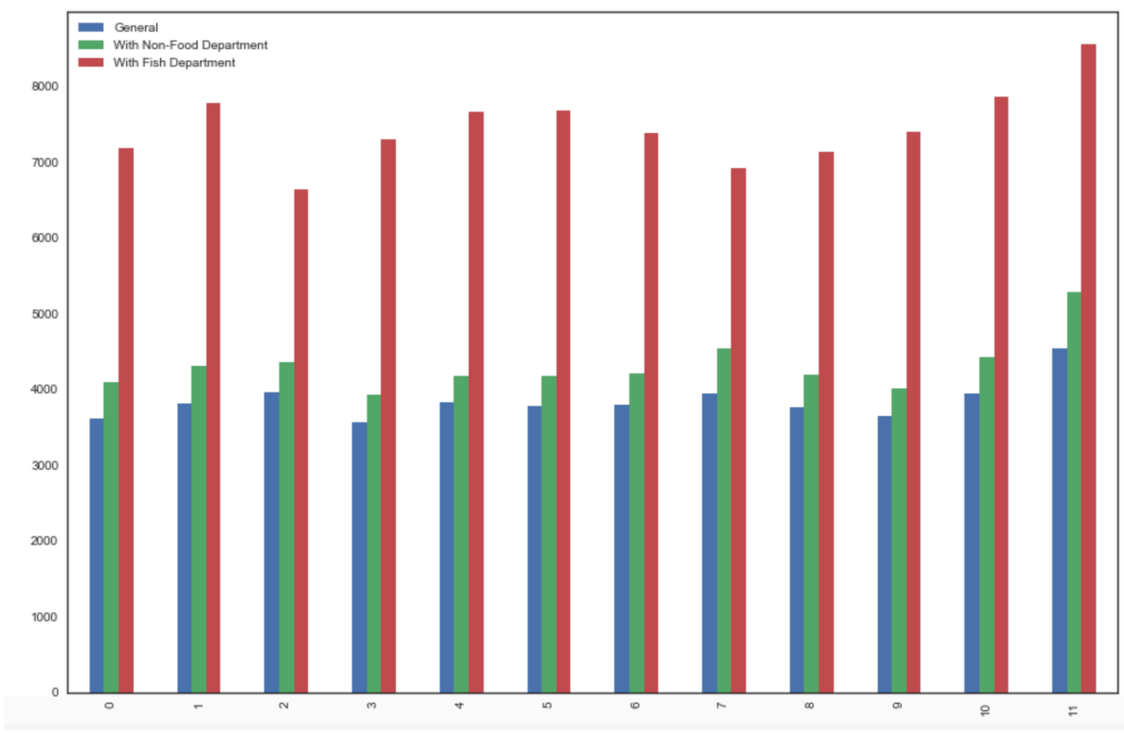
LORENZO NORCINI
GUGLIELMO MENCHETTI

DATA ANALYSIS

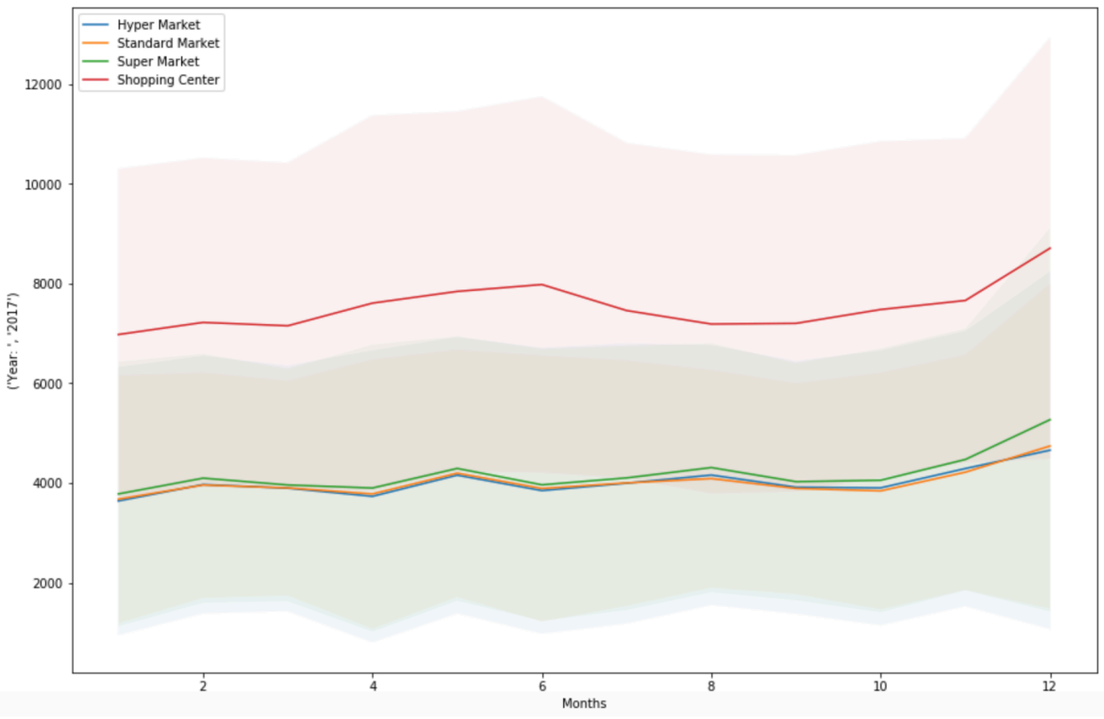
Monthly Sales per Region



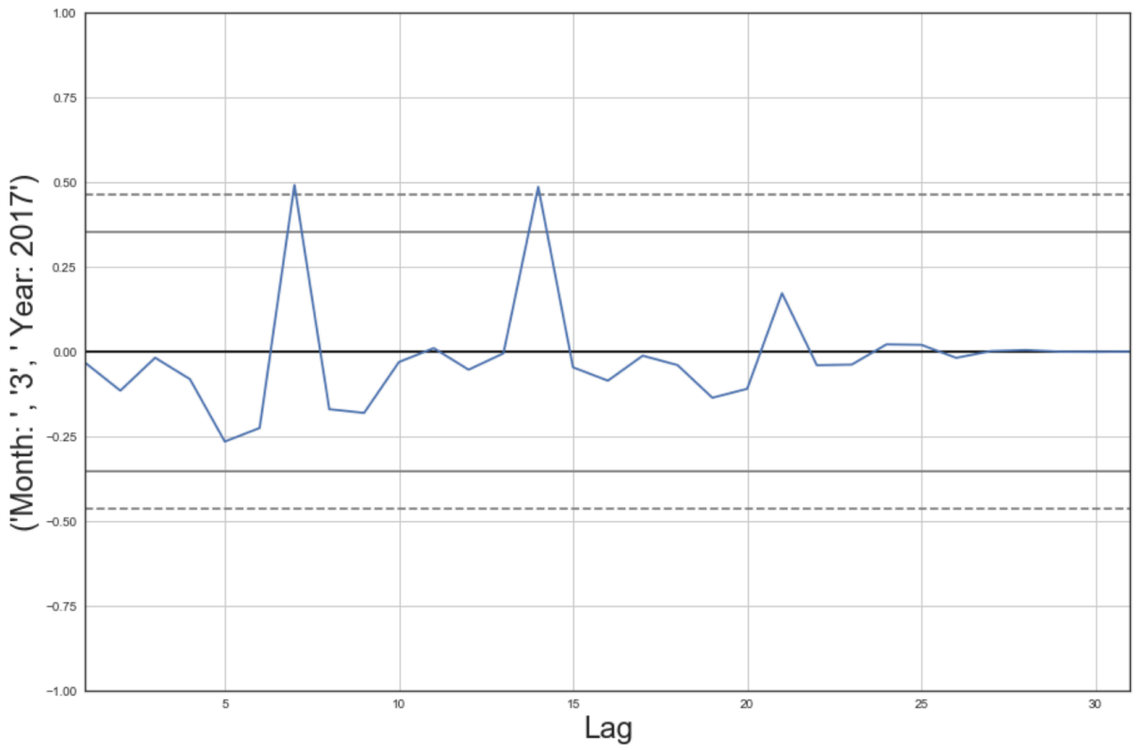
Monthly Sales per Assortment Type



Monthly Sales per Store Type



Autocorrelation

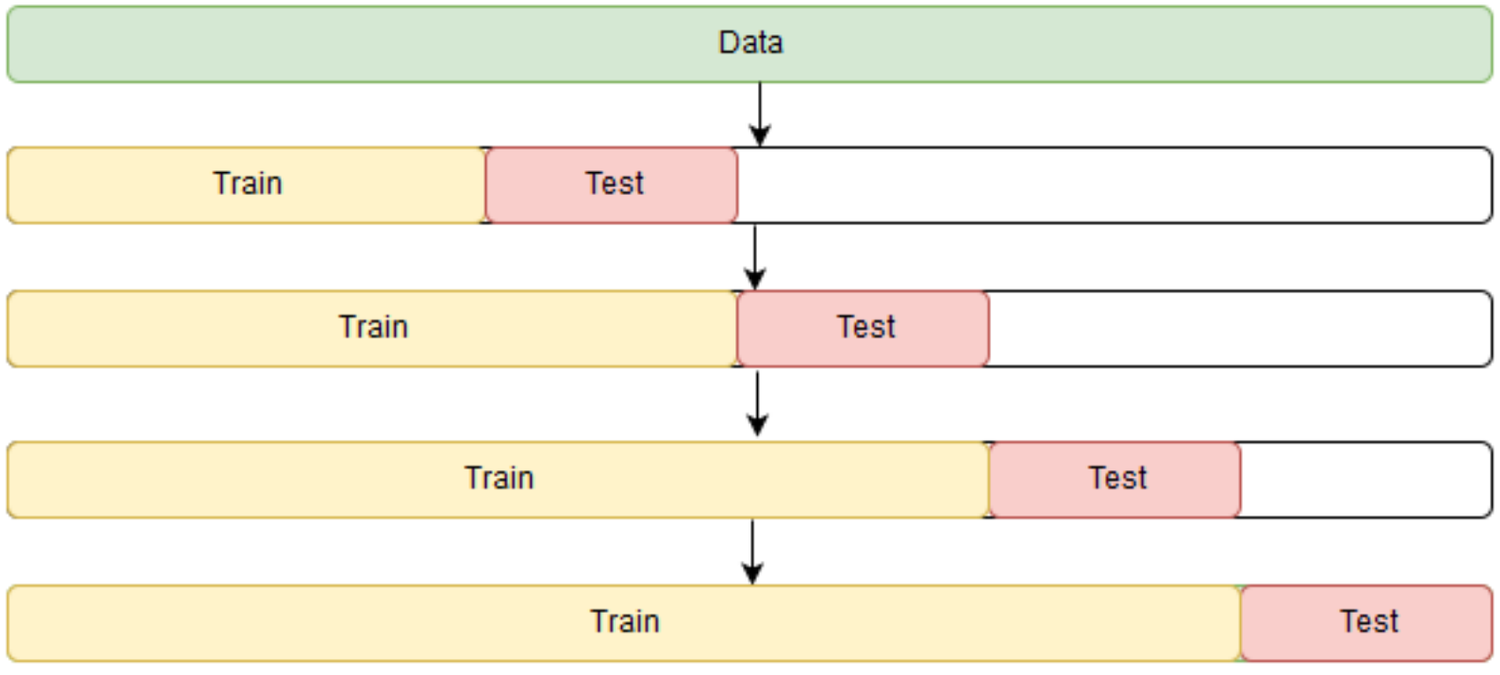


FEATURES AND MODEL PARAMETERS

Store Features	Time Features	Meteo Features
• IsHoliday • HasPromotions • StoreType • NearestCompetitor • Region • AssortmentType	• Rolling Mean (14, 30, 60, 90) • Lag (1, 7, 14, 21, 28) • Day of the week • Month	• Events • Cloud Cover • Precipitation

Model	Parameters
SVM	• Error Penalty (0.1, 1, 10, 100) • Kernel (rbf, linear)
Random Forest	• Number of Estimators (10, 20, 50, 100) • Bootstrap (True, False) • Max Features (all, log2, sqrt) • Min samples for leaf (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
K Nearest Neighbours	• Number of Neighbours (5, 10, 20, 50, 75, 100) • Distance Metric (Euclidian, Chebyshev, Manhattan) • Weights (Uniform, Distance)
Adaboost + KNN	• Number of Estimators (10, 20, 50, 100) • Number of Neighbours (5, 10, 20, 50, 75, 100) • Distance Metric (Euclidian, Chebyshev, Manhattan) • Weights (Uniform, Distance)

FEATURES AND MODEL SELECTION



Time Series
Cross Validation

2 Months Overlapped
Test set size

Performance Metric

$$E_r = \frac{\sum_{i \in S_r} \sum_{j \in \{3,4\}} |a_{ij} - p_{ij}|}{\sum_{i \in S_r}^{i} \sum_{j \in \{3,4\}}^{j} a_{ij}}$$ ***Region Error***

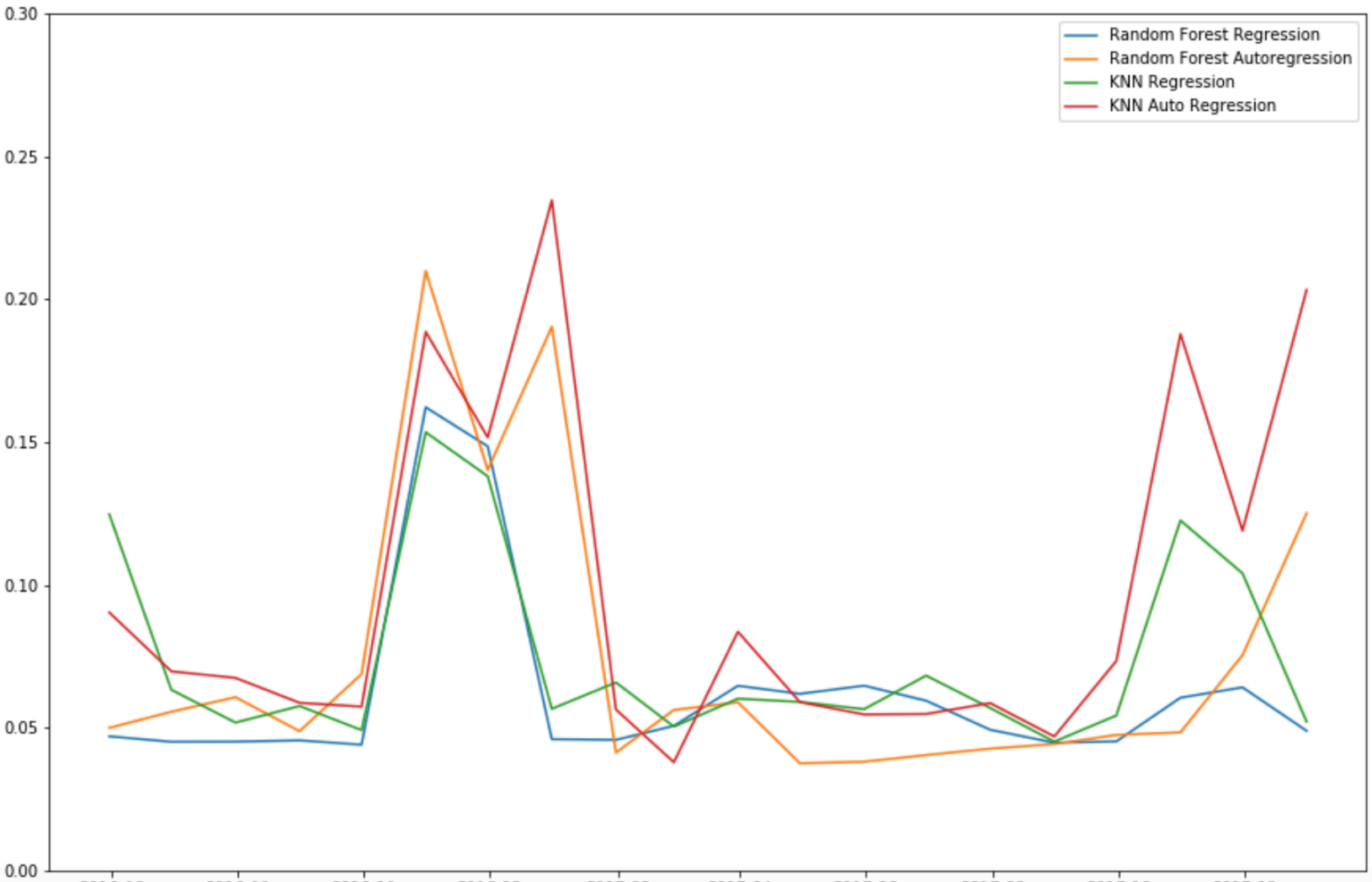
$$E = \frac{\sum_{r \in R} E_r}{|R|}$$ ***Total Error***

R *Set of regions*

a_{ij} *Actual Value of Store i and month j* p_{ij} *Predicted Value of Store i and month j*

PERFORMANCES

Type	Learner	μ	σ
Regression	K-NN	0.087	0.05
Regression	Random Forest	0.062	0.03
Autoregression	K-NN	0.097	0.06
Autoregression	Random Forest	0.074	0.05



Performances on Test folds

Learning Curves

OBTAINED MODELS

	Standard Regression	Auto Regression
Features	• IsHoliday • HasPromotions • StoreType • NearestCompetitor • Day of the week • Region • Month • AssortmentType	• IsHoliday • HasPromotions • StoreType • NearestCompetitor • WeekDay 'Region, • Month • AssortmentType • Lag 7 Days • Lag 14 Days • Rolling mean 60 Days
Model	Random Forest	Random Forest
Parameters	• Number of Estimator : 100 • Bootstrap • Max Features : all • Min samples for leaf :5	• Number of Estimator : 50 • Bootstrap • Max Features : all • Min samples for leaf :1
μ	0.062	0.074
σ	0.03	0.05

DATA MINING AND TEXT MINING

THANK YOU FOR YOUR ATTENTION

LORENZO NORCINI
GUGLIELMO MENCHETTI