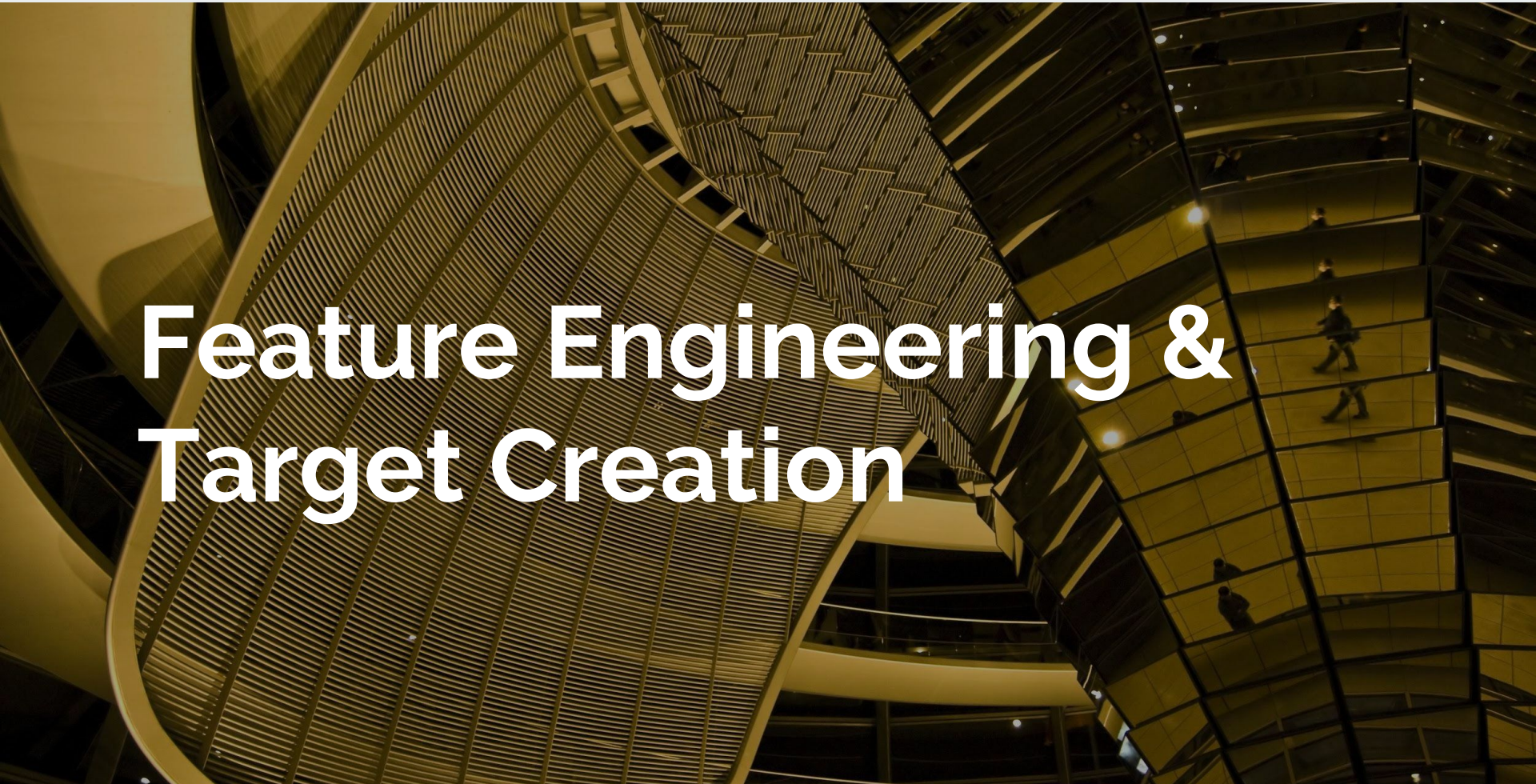
Surrounding Entities Impacting Sales

Identify the most important surrounding entities impacting POS

TOC


Feature Engineering &
Target Creation

Target Variable - Histogram Analysis (HA)

Classification Problem : (High Sales Shop = 1, Not High Sales = 0)

1) Separated data into train and split set 70:30 ratio
2) Created a histogram of log of sales values of the training set .
3) Based on visual inspection choose a cut-point, such that you have almost equal data-points on both sides
4) Binarized the target variable (train & test) based on the cut point.
5) Ensured that the cut-point is chosen, such that ratio of classes in the train and test data are similar .

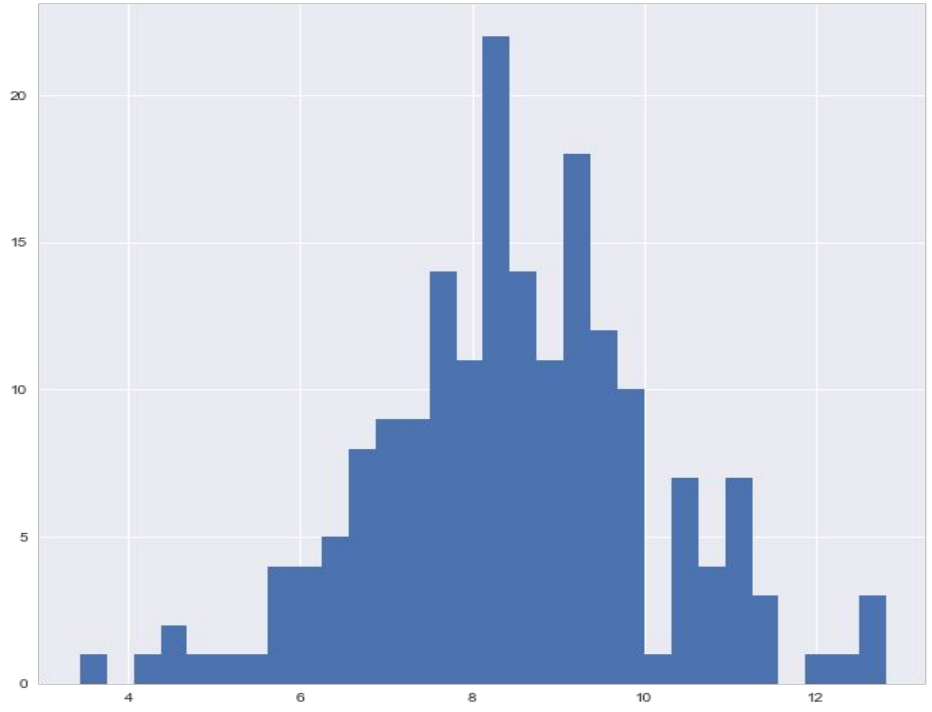
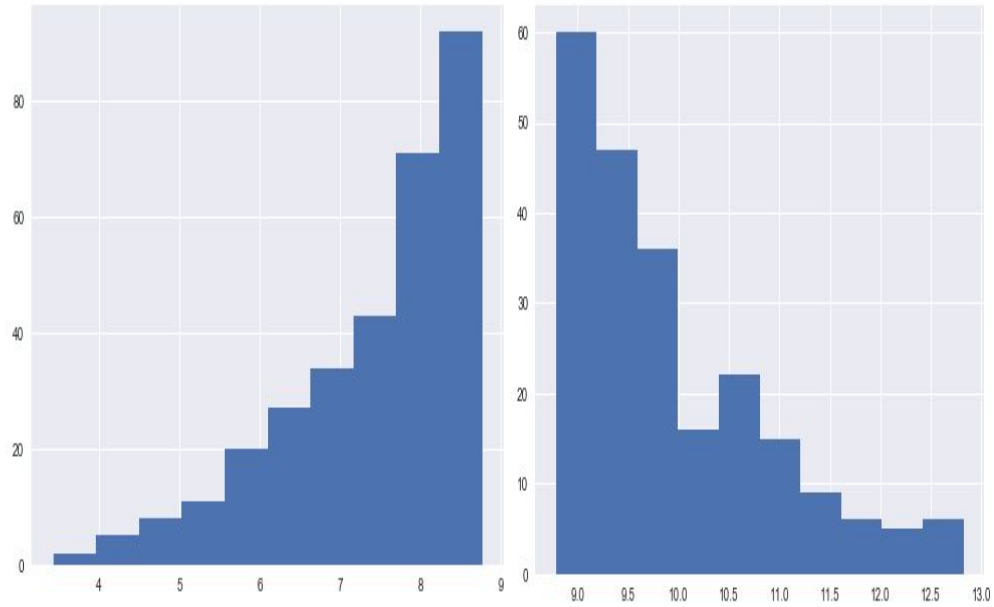


Fig : Log Total Sales

Target Variable - KMeans Clustering (KMC)

Classification Problem : (K=2)

1) Created new variable based on sales pattern of the store (Morning,Afternoon,Evening.Night)
2) Created two clusters (Kmeans K=2) based on the above variables and total sales information .
3) Checked if the total sales values of these clusters do not overlap



Log Total Sales , Cluster Label = 0

Log Total Sales, Cluster Label = 1

Creation of Features

Assumptions :

No.of ratings is a proxy for no.of people visiting the place .

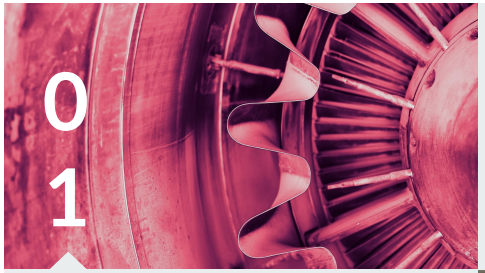
Overall rating indicates the attractiveness of the place (People would go/ return to places with high rating .

A few natural features already present , in the json document (a bit of parsing required).eg no.of entity_types around the store (bars, train station etc) . Following custom built features were created .

1. Average entity (per entity type) shared by each store .
2. Average rating of an entity (per entity type) shared by each store .
3. No.of stores around a particular store calculated on post-code level .ie stores in the same surrounding .
4. Average no.of reviews of an entity(per entity id) shared by each store .


Experiment
Design

Experiment Design



Feature Selection

Of all the features created, a few of them were selected based on
Logistic Regression with L1 penalty
Select KBest (f_classify methods)

Interpretation

All the classifiers used ,give us an indication of feature importance . We try to compare the output of various classifiers and select the most common features..

Data Cleaning
Target & Feature Creation

Targets were created in two different ways and various features were created too, focussing on interaction between store and entity on a micro level .

Classification

The features selected in the feature selection phase were passed on to DT,RF,ET classifiers. The results of the classification give us an confidence , of the experiment setup .

Getting Ready- Common Settings

Training & Testing datasets

01

Test data ratio : 0.3

Model Evaluation

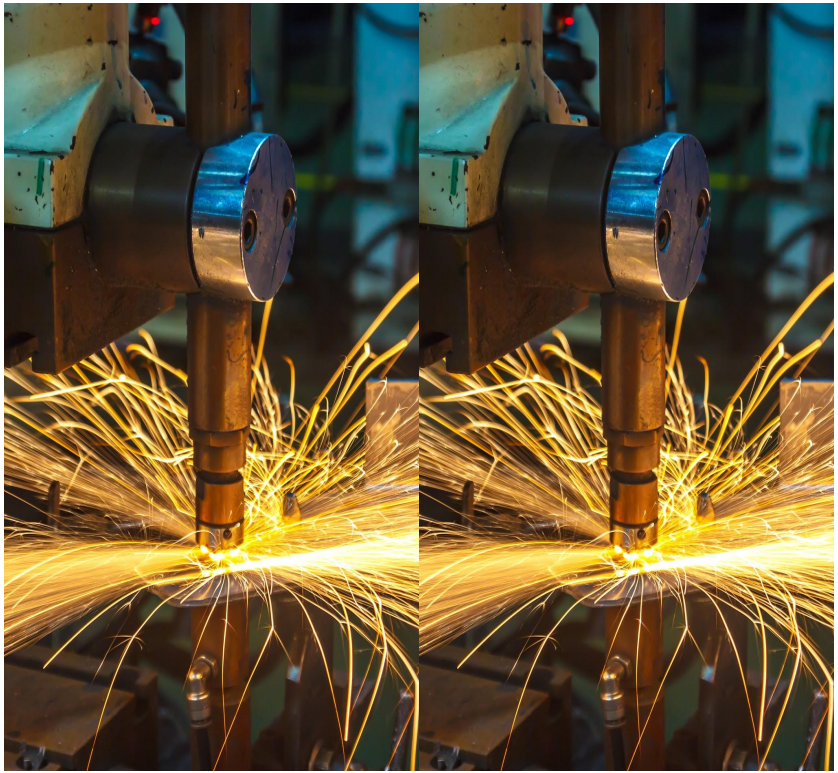
02

K Fold cross validation (k=10)
ROC-AUC method .

Grid Search

03

RandomizedSearchCV



Feature Selection

Logistic Regression with L1 Penalty

01

Logistic regression with L1 penalty was used to preselect features .

Benefit of L1 is that it can push feature coefficients to 0, creating a method for feature selection.

Training / Testing Scores : 0.80 /0.70

Feature Selection

Select KBest Features

02

Scikit learn library provides out of the box methods for feature selection . eg Kbestfeatures

Compute the ANOVA F-value between each feature and the target.

Use the selected features as input to other classifiers.

Classification

Decision Tree

01

Though Decision tree classifiers suffer from the disadvantages such as growing complex and still unable to generalise the data well.

It is simple to understand and trees can be visualised .

Training / Testing Scores : 0.61 / 0.60

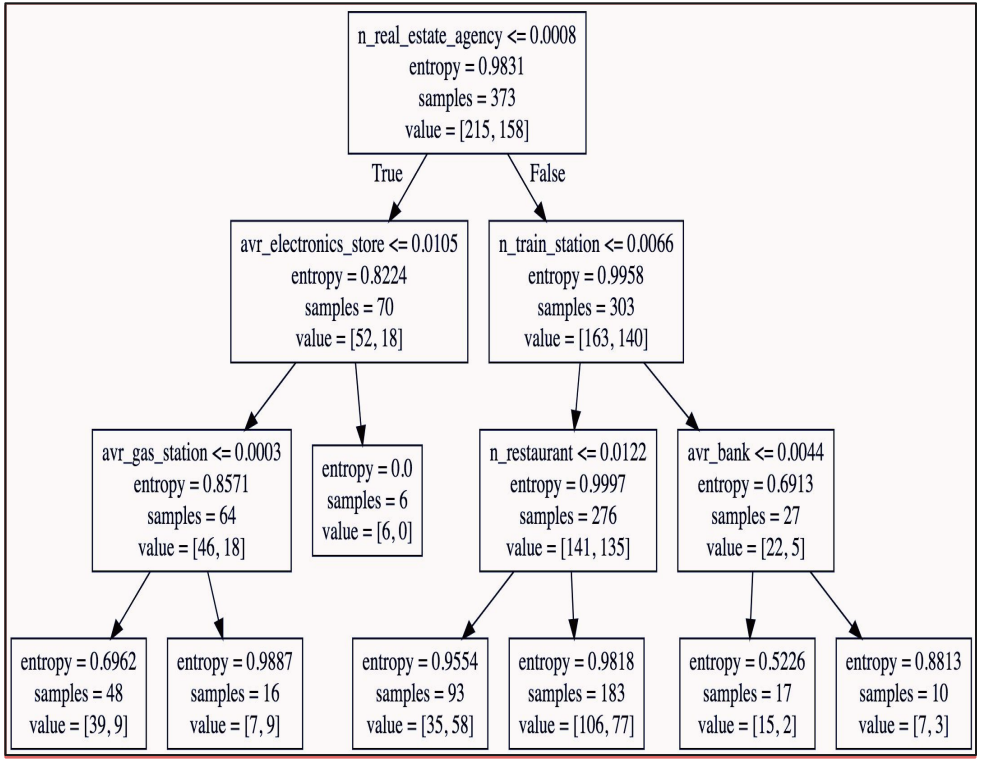


Fig : Decision Tree built on Target created by KMeans clusterting

Classification

Random Forest & Extra Trees Classifier

02

One of the most robust set of classifiers, inherently reduces overfitting .

Provides a list of feature importance, which can help in identifying entity attributes.

Random Forest :
Training / Testing scores : 0.75/0.68

Extra Trees Classifier :
Training / Testing scores : 0.80/0.75

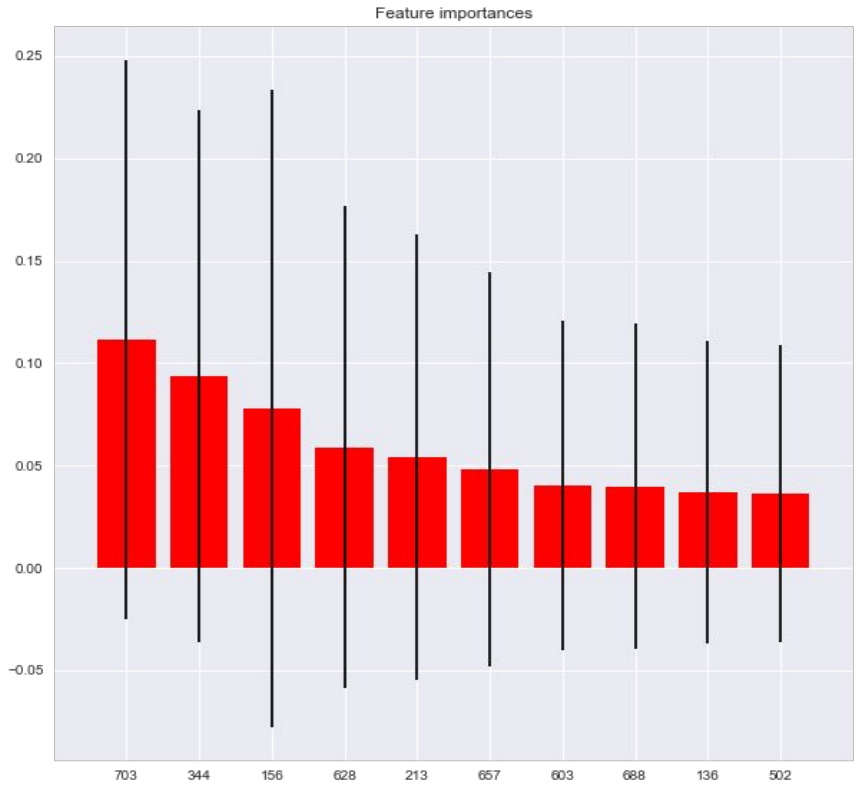


Fig : Extra Tress Classifier Feature Importance


Results

Performance Analysis

A run is considered good enough for usage if :

- Train & Test ROC-AUC is >= 0.60
- Difference between Training AUC and Testing AUC <= 10 percentage points

Classifier	Decision Tree Clf (1)	Extra Tree Clf (2)	Decision Tree Clf (3)	Random Forest Clf (4)	Extra Tree Clf (5)	Decision Tree Clf (6)	Random Forest Clf (7)
Target creation Method	HA	HA	HA	HA	HA	KMC	KMC
Feature Preselection Method	NA	NA	KBest(F_classif)	KBest(F_classif)	Logistic Regression(L1 Penalty)	NA	Logistic Regression(L1 Penalty)
Training AUC	0.74	0.74	0.68	0.68	0.69	0.72	0.68
Test AUC	0.69	0.64	0.61	0.63	0.61	0.64	0.62

Entities Affecting Sales

- Features related to the same entity type have the same color code.

(1)Decision Tree Clf NA HA (0.74/0.69)	(3)Decision Tree Clf KBest(F_classif) HA (0.68/0.61)	(6)Decision Tree Clf NA KMC (0.72/0.64)	(2)Extra Tree Clf NA HA (0.74/0.64)	(4)Random Forest Clf KBest(F_classif) HA (0.68/0.63)	(5)Extra Tree Clf Logistic Regression(L1 Penalty) HA (0.69/0.61)	(7)Random Forest Clf Logistic Regression(L1 Penalty) KMC (0.68/0.62)
n_rating_per_entity_store_beauty_salon	total_user_rating_doctor	avr_lodging	rating_per_entity_store_physiotherapist (0.087518)	rating_per_entity_store_hair_care (0.112934)	rating_per_entity_store_hair_care (0.056445)	n_review (0.161625)
entity_per_store_real_estate_agency	total_user_rating_beauty_salon	n_hardware_store	city_GR (0.071863)	total_user_rating_beauty_salon (0.110141)	rating_per_entity_store_insurance_agency (0.054403)	n_doctor (0.119059)
total_user_rating_gas_station		n_pharmacy	rating_dentist (0.060138)	total_user_rating_hair_care (0.103838)	n_rating_per_entity_store_pharmacy (0.035163)	n_real_estate_agency (0.070792)
n_beauty_salon		city_VD	entity_per_store_jewelry_store (0.054807)	total_user_rating_doctor (0.09116)	n_rating_per_entity_store_dentist (0.034688)	avr_gym (0.070436)
n_rating_per_entity_store_real_estate_agency		avr_atm	rating_home_goods_store (0.054117)	n_rating_per_entity_store_doctor (0.079437)	n_store (0.031897)	avr_insurance_agency (0.062273)
		n_train_station	entity_per_store_bar (0.050996)	n_review (0.079117)	rating_per_entity_store_spa (0.031538)	n_store (0.053597)
			avr_gym (0.040799)	n_gym (0.064840)	avr_post_office (0.031239)	avr_cafe (0.044384)
			avr_electrician (0.039407)	rating_per_entity_store_dentist (0.062976)	n_rating_per_entity_store_gym (0.028741)	avr_spa (0.039855)
			n_store (0.038358)	n_store (0.059283)	rating_per_entity_store_furniture_store (0.027427)	n_clothing_store (0.039071)

Thank you.