Predicting Success of Restaurants

Under the guidance of Dr. Dinesh Singh

GROUP MEMBERS

- Shivam Anand, 20164050
- Satyam Singhal, 20164047
- Somesh Gupta, 20164161
- Uddesh S. K. Singh, 20164141

MOTIVATION

The well being of many businesses today heavily rely on the positive ratings given by their customers and are being data driven. With the founding of Yelp in 2004, the relationship between businesses and their customers has become more dynamic and we have more access to business data. In this project, we studied the success of restaurants by predicting the star ratings of restaurants and finding the most useful traits in determining their success.

HOW CAN WE USE MACHINE LEARNING TO PREDICT THE SUCCESS OF A BUSINESS ?

We worked on the following two criteria to determine the results

- **Business attributes and finding the best attributes for success.**
- **Sentiment analysis on textual reviews.**

WORKING ON BUSINESS ATTRIBUTES

Our work on the first criterion can be represented broadly by

- **Dataset Gathering**
- **Preprocessing**
- **Training Data Models**
- **Result and Analysis**

DATASET GATHERING

- The dataset we worked upon was the Yelp Open Dataset.
- The Yelp dataset is a subset of the entire businesses, reviews, and user data available for use in personal, educational, and academic purposes.
- Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.
- It contains **6,685,900 reviews** of over 192,609 businesses spreading to 10 major cities, which were not limited to just food establishments.

DATA INFO

Data Instances(RAW json) - 3164, 15

City used for Prediction - Cleveland

Data Instances(initial preprocessing) - 3164, 737

Data instances(After removing fields with >80% null values) - 3164, 703

Data instances(choosing just restaurant businesses) - 1361, 703

PREPROCESSING

Due to the sheer amount of raw and unstructured information, we had to conduct a fair bit of preprocessing that took up most of the time.

- **Extraction of data**
  - City-wise: Data contains information of 10 major cities. We had to be localised. Thus only the data for Cleveland was extracted.
  - Establishment-wise: Data for only Food and Beverages establishments were extracted.
- **Extracting individual attributes**
  - Attribute column was a singular entity with string entries divided by commas, not helpful, each entry was individually extracted.
- **Enumerating extensive individual attribute columns**
  - Each extracted entity was enumerated individually in the training file as a boolean attribute.
- **Attributes existing in only minority establishments were excluded**
  - Attributes that existed for less than 20% of the establishments were excluded to increase efficiency.

DATA TRAINING MODELS

We used a variety of data training models to obtain optimised results

Regressor

- Linear Regression
- Decision tree regressor
- Support Vector Regressor

Classifier

- Logistic Regression Classifier
- Support Vector Classifier
- K Nearest Neighbours

LINEAR REGRESSION

linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

We used three different methods to fit training data and predict on testing data

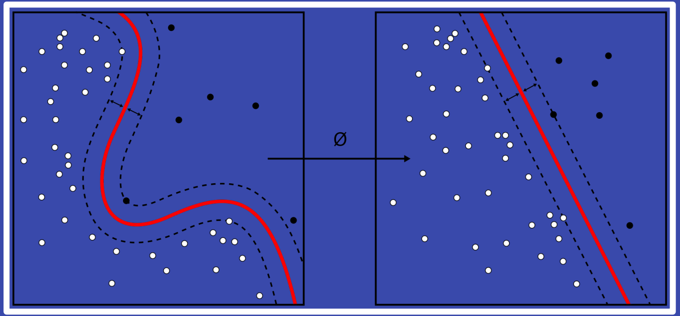
- Train Test split (66:33) - Split arrays or matrices into random train and test subsets
- Cross Validation
- Polynomial features

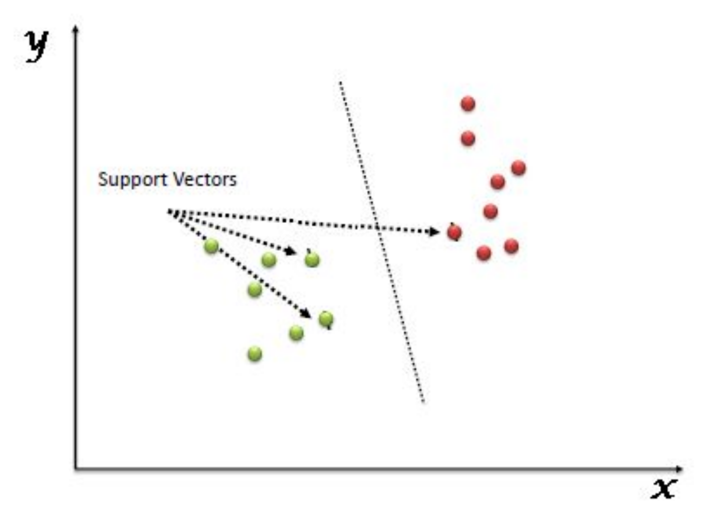
LOGISTIC REGRESSION

- In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).
- Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail, win/lose, alive/dead or healthy/sick; these are represented by an indicator variable, where the two values are labeled "0" and "1".
- In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).
- The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function

SUPPORT VECTOR MACHINE

Support-vector machines (**SVMs**, also **support-vector networks**) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.





The figure shows how Kernel machines are used to compute

Non-linearly functions into a higher dimension linearly separable function.

MULTINOMIAL NAIVE BAYES

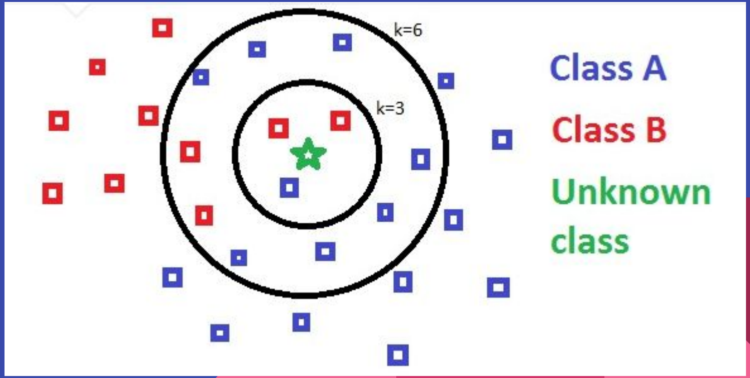
- The Naive Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naïve independence assumptions. It is one of the most basic text classification techniques.
- Multinomial Naive Bayes is used when the multiple occurrences of the words matter a lot in the classification problem. Such an example is when we try to perform Topic Classification
- It estimates the conditional probability of a particular word/term/token given a class as the relative frequency of term t in documents belonging to class c:

$$P\left(t \mid c\right) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

- Thus this variation takes into account the number of occurrences of term t in training documents from class c, including multiple occurrences.

K NEAREST NEIGHBOURS

- In pattern recognition, the ***k*-nearest neighbors algorithm** (***k*-NN**) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space
- *k*-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.
- The *k*-NN algorithm is among the simplest of all machine learning algorithms.



HYPERPARAMETERS APPLIED TO TRAIN REGRESSION MODELS

Hyperparameters	
Restaurant Characteristics	
normalize(LinearReg)	False
max_features(DTReg)	0.5
max_depth(DTReg)	4
C(SVR)	1.0
Gamma(SVR)	0.01

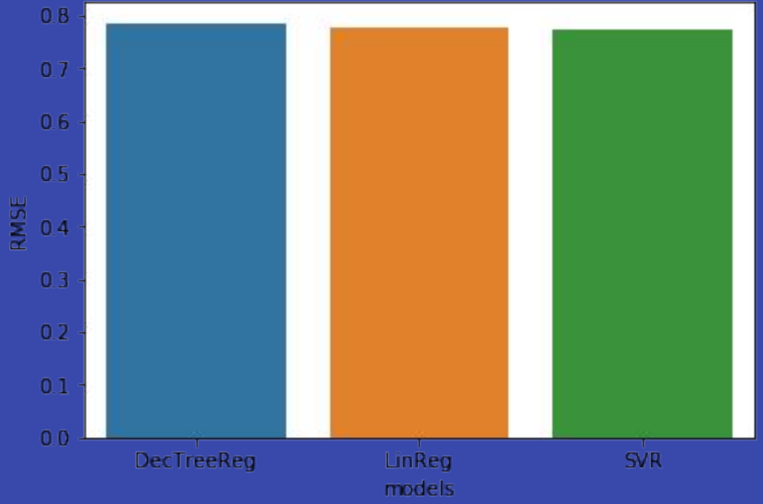
HYPERPARAMETERS APPLIED TO TRAIN CLASSIFICATION MODELS

Hyperparameters	
Restaurant Characteristics	
n_neighbours(KNN)	3
n_jobs(KNN)	-1
weights(KNN)	distance
leaf_size(KNN)	2
algorithm(KNN)	ball_tree

Hyperparameters	
Restaurant Characteristics	
kernel(SVC)	rbf
C(SVC)	6
penalty(LogReg)	l2
C(LogReg)	0.001

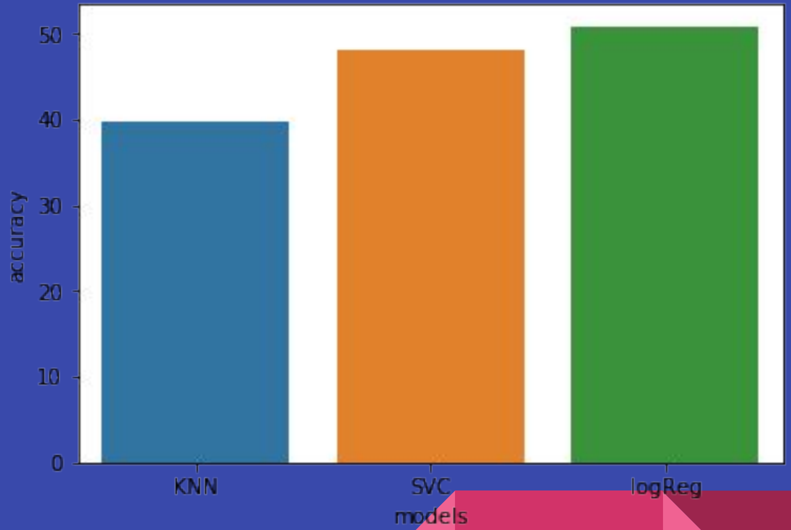
RESULTS

Results	
Restaurant Characteristics	
MODEL	**RMSE**
Linear Regression	0.7765
Dec. Tree Regression	0.7852
Support Vector Regression	0.7743



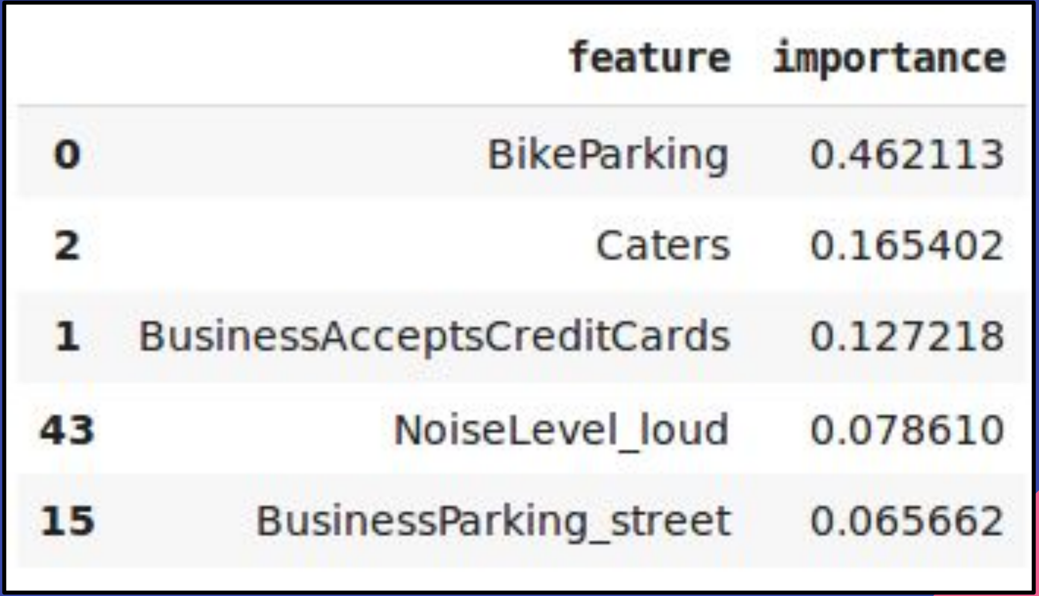
RESULTS(Contd.)

Results	
Restaurant Characteristics	
MODEL	**ACCURACY**
Logistic Regression Classifier	50.918
Support Vector Classifier	48.052
K Nearest Neighbour	39.676



RESULTS(Contd.)

Attributes with the most importance and weightage



	feature	importance
0	BikeParking	0.462113
2	Caters	0.165402
1	BusinessAcceptsCreditCards	0.127218
43	NoiseLevel_loud	0.078610
15	BusinessParking_street	0.065662

WORKING ON SENTIMENT ANALYSIS OF TEXTUAL REVIEWS

Our second criterion for determining our predictions were Textual reviews and our task could be main classified in to the following broad categories

- **Data Gathering**
- **Preprocessing**
- **Training Models**
- **Result Analysis**

PREPROCESSING

- **Extraction of star rating and textual reviews**
  - Extracting star rating and textual reviews of users from the chosen data of Food establishments, for the city of cleveland.
- **Grouping of data according to establishments**
  - All reviews for a particular establishment were weighed according to users and their helpfulness and merged into a single field, resulting in one overall review per establishment.
- **Balancing of dataset to remove bias**
  - The reviews were classified according to star rating and the rating with minimal reviews was chosen as a baseline to keep the amount of reviews equal.
- **Vectorizing and feature creation**
  - Each review is then tokenized into unigrams and bigrams, followed by vectorization and polarization of adjectives using tf-idf.

PREPROCESSING(Contd.)

- **Tokenization** -The process of segmenting running text into words and sentences.
- **Vectorization** - vectorization of a matrix is a linear transformation which converts the matrix into a column vector.
- **Stopwords** - In computing, stop words are words which are filtered out before or after processing of natural language data (text). Though "stop words" usually refers to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list.
- **TF-IDF** - TF-IDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
  - The weight of a term that occurs in a document is simply proportional to the term frequency.
  - The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.
- **N-grams** - n-gram models are widely used in statistical natural language processing.For parsing, words are modeled such that each n-gram is composed of n words. We here, have worked with unigrams and bigrams.

DATA INFO

Data Instances(RAW json) - 91654, 2

City used for Prediction - Cleveland

Data Instances(used) - 91057, 2

Data Instances(of minimal star rating) - 8074

Data Instances(upon balancing) - 40370, 2

Tokens used after vectorization - Top 500, most frequent tokens

Data Instances(trained upon) - 40370, 500

DATA TRAINING MODELS

We used a variety of data training models to obtain optimised results

Regressor

- Linear Regression
- Decision tree regressor
- Support Vector Regressor
- Multinomial Naive Bayes

Classifier

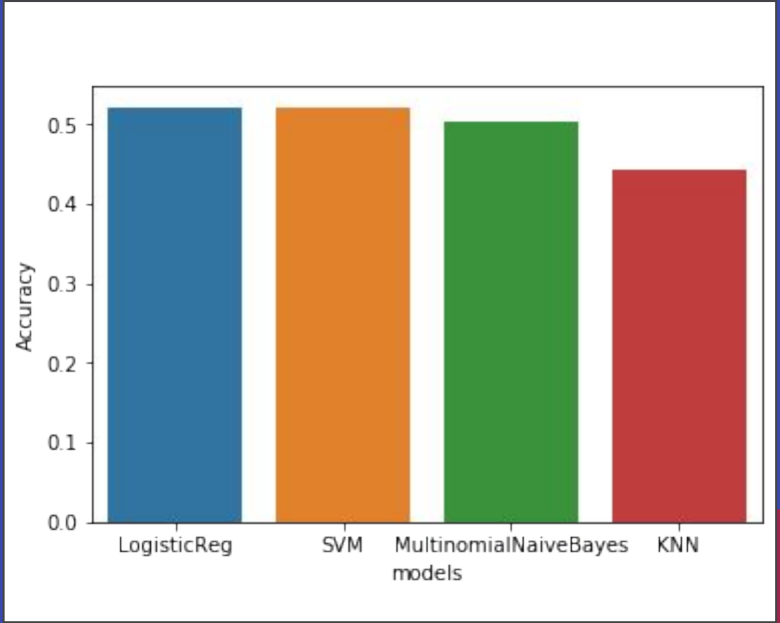
- Logistic Regression Classifier
- Support Vector Classifier
- K Nearest Neighbours

HYPERPARAMETERS APPLIED TO TRAIN REGRESSION MODELS

Hyperparameters	
Restaurant Review Texts	
fit_intercept(LinearReg)	True
normalize(LinearReg)	False
max_features(DTReg)	0.5
max_depth(DTReg)	8
C(SVR)	1000.0
Gamma(SVR)	1.0

RESULTS

Results	
Restaurant Review Texts	
MODEL	**ACCURACY**
Support Vector Classifier	52.075
K Nearest Neighbours	44.240
Logistic Regression	52.060
Multinomial Naive Bayes	50.298

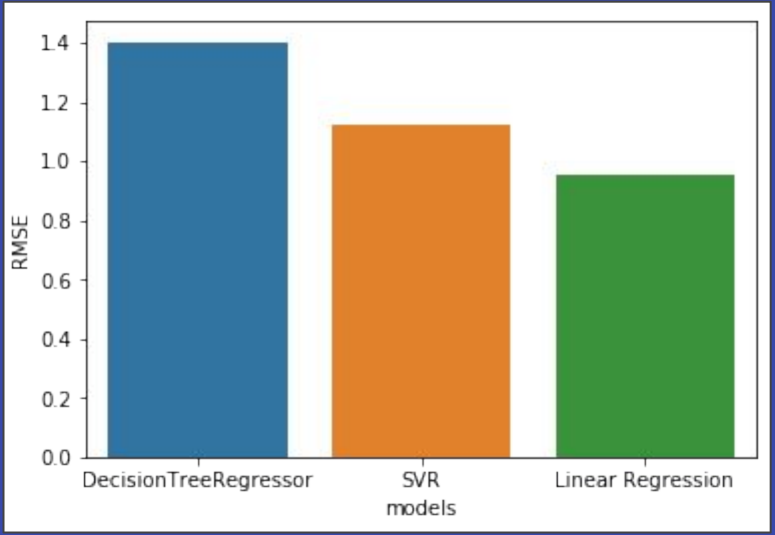


HYPERPARAMETERS APPLIED TO TRAIN CLASSIFICATION MODELS

Hyperparameters	
Restaurant Review Texts	
n_neighbours(KNN)	3
leaf_size(KNN)	2
Penalty(LogisticReg)	l1
C(LogisticReg)	1.0
C(SVC)	100.0

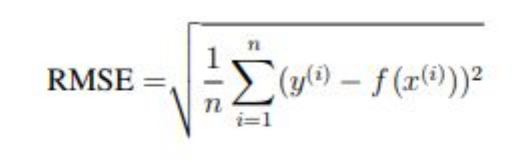
RESULTS(Contd.)

Results	
Restaurant Review Texts	
MODEL	**RMSE**
Linear Regression	0.9514
Dec. Tree Regression	1.4023
Support Vector Regression	1.1194



RESULT METRIC

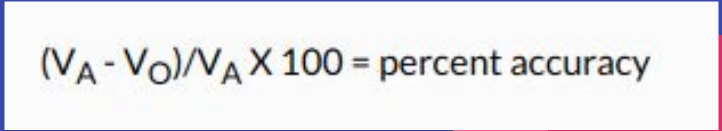
RMSE - Root Mean Square Error.



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - f(x^{(i)}))^2}$$

It is the square root of the variance in the predicted values when compared to the actual values of test cases, i.e., the deviation.

ACCURACY - Instances predicted correct over total Instances



$(V_A - V_O)/V_A$ X 100 = percent accuracy

CONCLUSION

- **Support Vector Machine** and **Logistic Regression** model give the most accurate result among classifiers, whereas **Linear Regression** and **Support Vector Regressor** give most close results among regressors ,with regards to our chosen dataset.
- We found least deviated result with a **RMSE of 0.7742 for regression** and best **ACCURACY of around 53% for classification**.
- The accuracy score of 53% suggest our models perform satisfactory, as on a random guess the probability of a correct classification will be 20%.
- If only sentiment of a review is determined, the accuracy score shoots up to 96% for the same dataset.

ANALYSIS

- As we see from the results of both cases we worked upon, **Support Vector Machine perform best among all the models.**
  - We can attribute this to the fact that the number of features(attributes) and instances of training data are of comparable magnitude and SVM works better in higher dimension.
  - The sparsity in our chosen data being high, also attributes to SVM being an optimum fit.
- We see that our classifier models are more suited for textual review analysis, and give a better accuracy score for the same.
- Regression gives better result on attributes dataset, due to the relatively less features there ,thus enabling correlation between features easy.
- Increasing the dataset significantly improves the result.

FUTURE WORKS

- Retraining the data and rerunning the tests on larger datasets, upon acquisition of more powerful hardware.
- Use of unsupervised learning algorithms in conjunction to the supervised learning algorithms.
- Using various more localised datasets in order to generate a heat map of popular cuisines and spending power of the populace.
- Using neural network to improve the accuracy of prediction.

REFERENCES

- Public data: http://www.yelp.com/dataset challenge
- Wang, Junyi. ”Predicting Yelp Star Ratings Based on Text Analysis of User Reviews.”
- Asghar Nabiha, ”Yelp Dataset Challenge: Review Rating Prediction.” 2016
- Vo, Kang, “Predicting Success of Restaurants in Las Vegas”, 2016

THANK YOU