

# Restaurant Sales & Status Prediction

【Machine Learning】 【Prediction】 【Algorithms】

May 2021



# Chapter 1 - Intro



There are numerous locations in Hong Kong where a restaurant owner could consider starting a new business. People who want to open a new restaurant but don't know anything about the area, district, or other variables will have a more difficult time planning and budgeting for it than people who want to open a restaurant based on data and evidence. We will use restaurant point of sale (POS) data, public data from Hong Kong government, and property agency to prepare this research and determine the best approach for opening a restaurant in Hong Kong.



**Accelerate the decision-making and transaction-closing processes:**  
We use machine learning to uncover hidden patterns and deliver location insights and recommendations by integrating diverse data sources.

# Problem Statement



01

To get a new restaurant site up and running, a significant amount of time and investment is required. When the wrong location is chosen for a restaurant brand, the site closes within 12-18 months and incurs operating losses.

02

There are over 5000 locations for which revenue need to be predicted based on various data fields mentioned in the dataset.

03

The main goal of this research is to predict a selection of locations that can survive and the revenue of the restaurants in the given dataset using machine learning algorithms from previously established restaurant data.

# Chapter 2 - Literature Review



Point out related researches and then examine the theories that underpin the Problem Statement



Select feature set, target and method algorithms – SVM, KNN, Radom forest regression, ANN, etc...



## Why Use Machine Learning?

- ML algorithm can simplify code and perform better.
  - ML can find complex problems
  - ML can easy adapt new data
- Easy to get insights of large amount of data

# Chapter 3 - Research Methodology



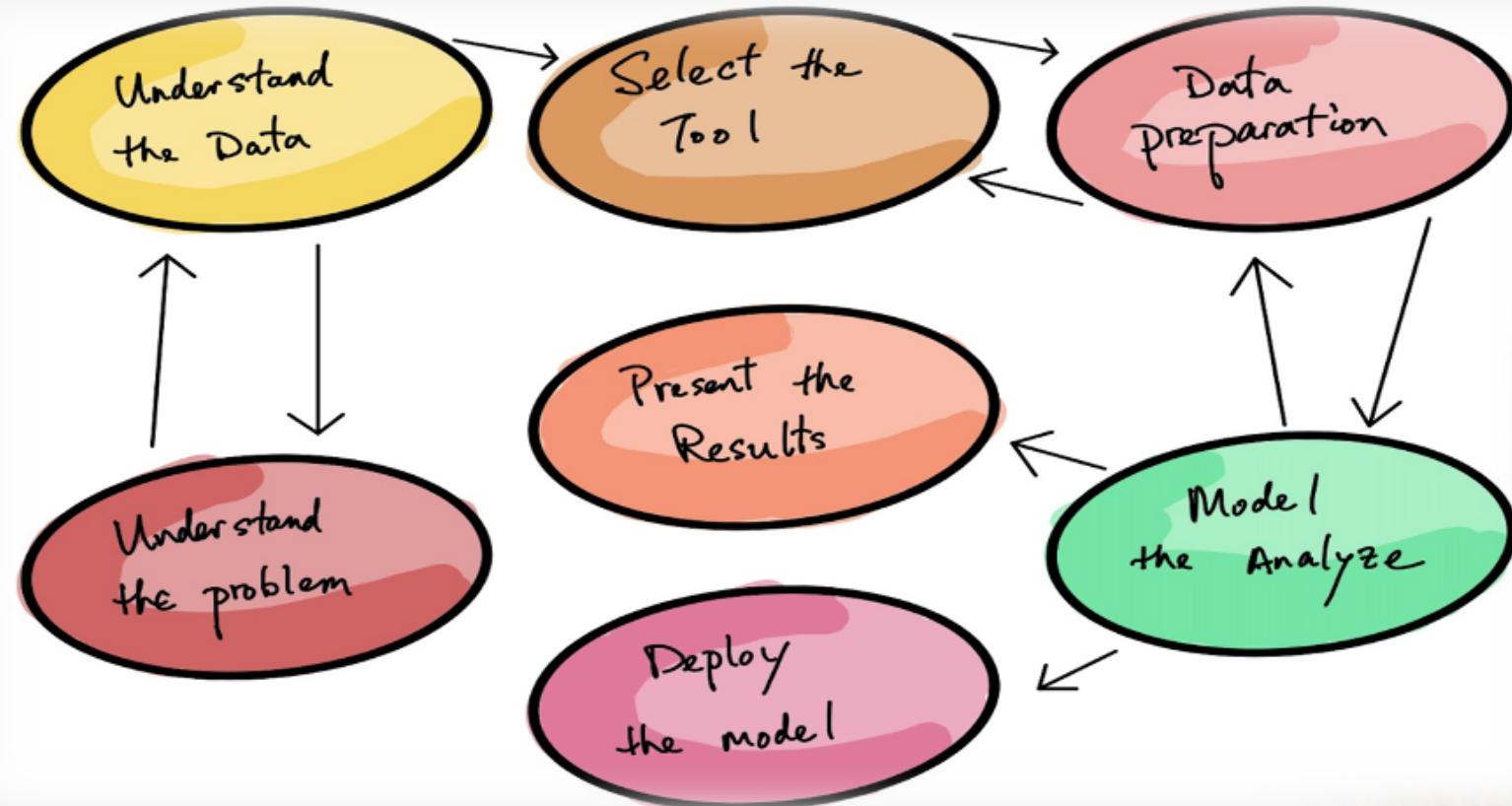
Based on Machine Learning Model, define the lifecycle of data analysis process from raw data to structured data, model building, evaluation and parameters tunings, visualization and model deployment.



## Data Science Analysis / Exploratory Data Analysis (EDA)

In a work offered the lifecycle of data analysis and development. We often want a numerical prediction over a categorical target. In current Restaurant Status research, a basic yes/no prediction of whether a restaurant is likely to continue for keep the business running may not be sufficient; We want to model the probability that the restaurant will continue. This is still considered classification modeling rather than regression because the underlying target is categorical.

# Exploratory Data Analysis (EDA)



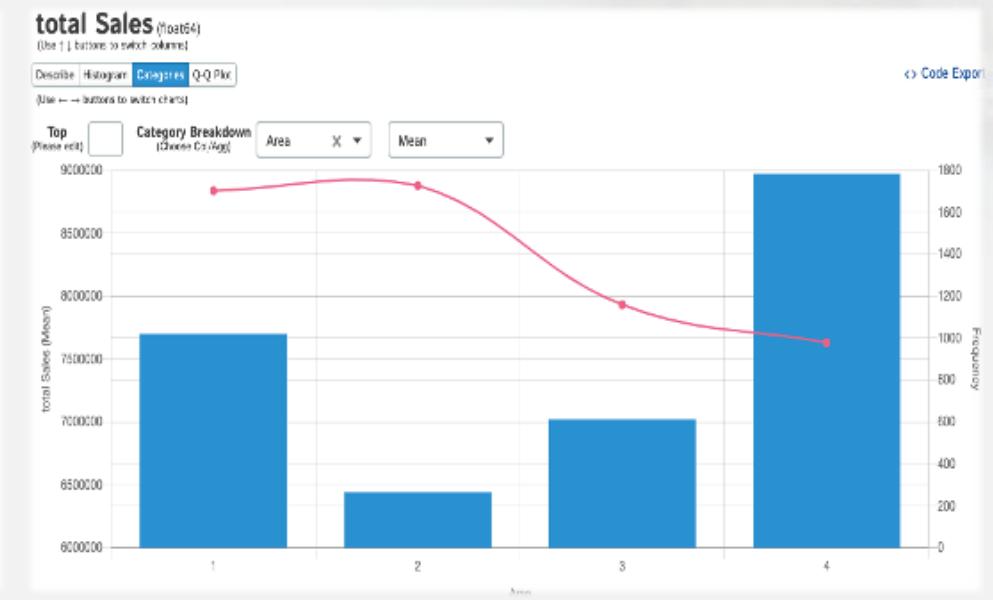
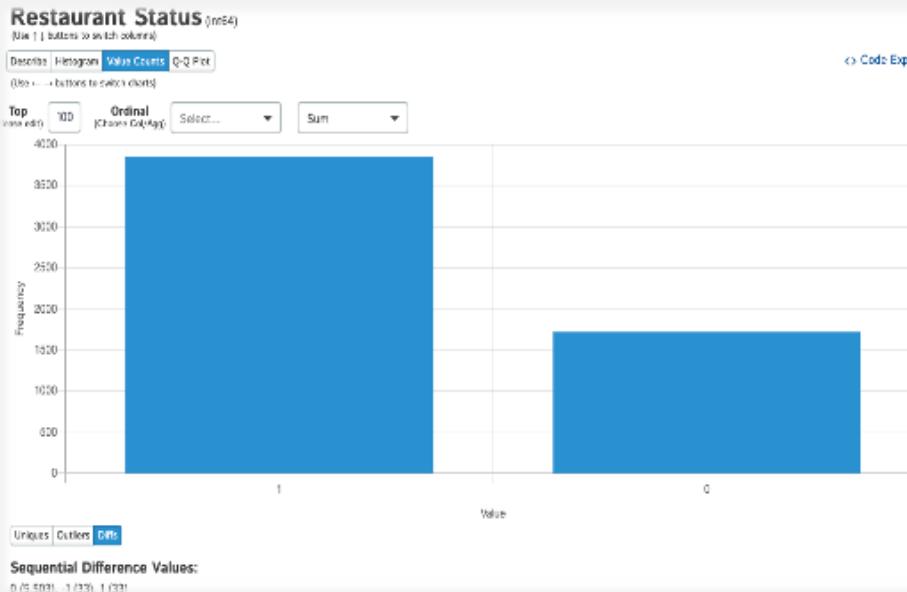
## Dataset consists of following fields:

- Restaurant id: ID
- Restaurant type: Eg. Chinese, western, fast food, etc..
- Area:
- District:
- Start Day, month, year
- District – Number of People able to work 15 to 65+
- Area – Number of worker 15 to 65+
- Total Sales (revenue)
- Payment type(cash,visa/master, others)
- Number of transaction
- Restaurant Sqf
- Monthly rent
- Location type (eg. shop in the street, mall, school, etc...)
- Restaurant Status ( Closed or Open )



# Insight of Dataset

- There are 5570 rows and 41 features in training data
- Training data shows variation in the “status” correlation with district who are able to work and “total sales” attribute related to the payment type.
- Area – Kowloon, Kowloon East etc.. District – TST, MK, WC etc.. Type of restaurant – Fast food, Chinese Cuisine, Japanese etc... Location type – Mall, Street, building etc...



# This research of tools in Python

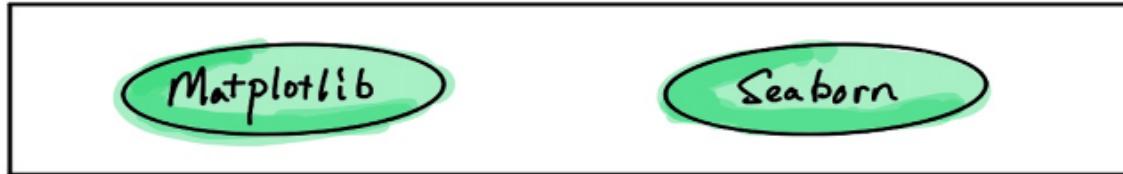
Data Representation



Machine Learning Statistical



Data Visualisation



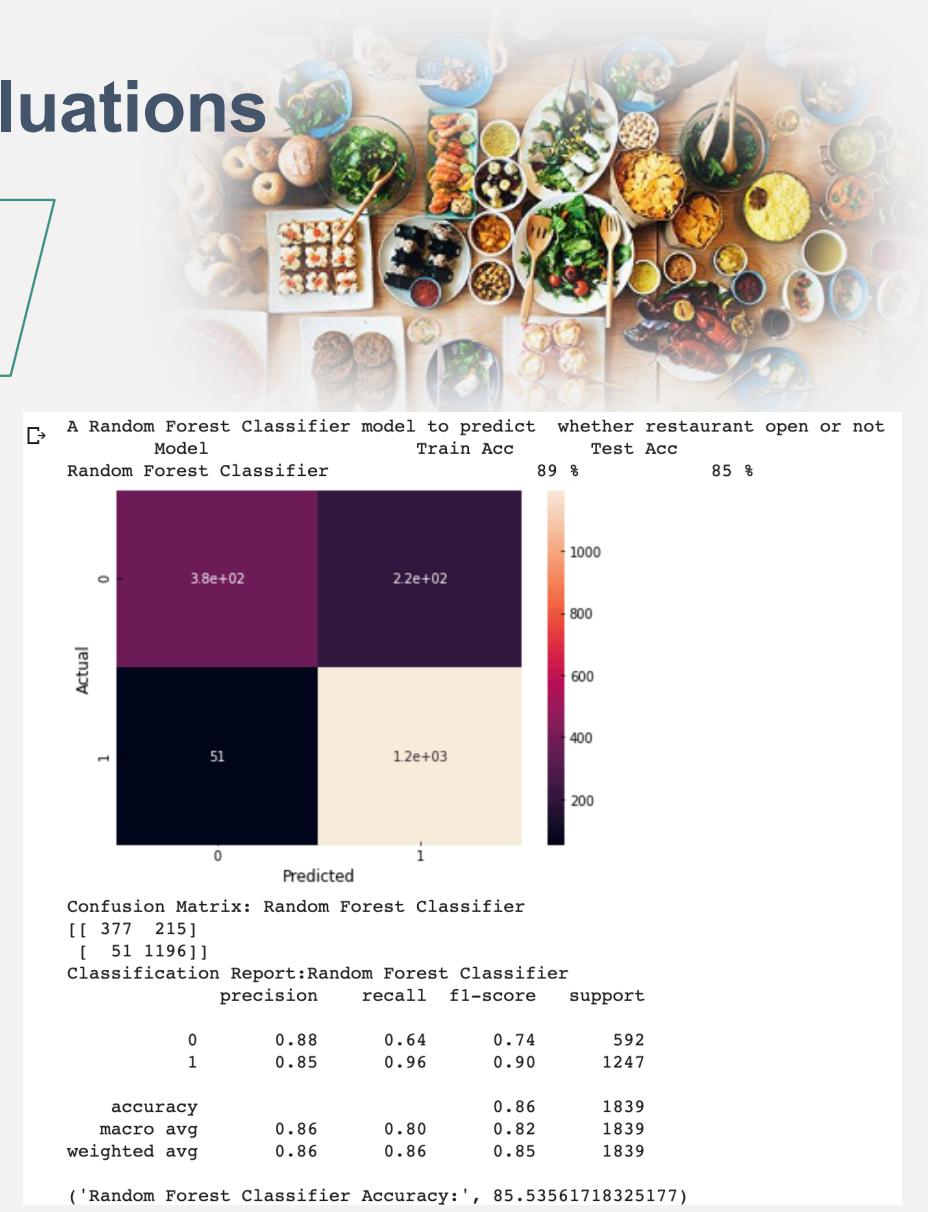
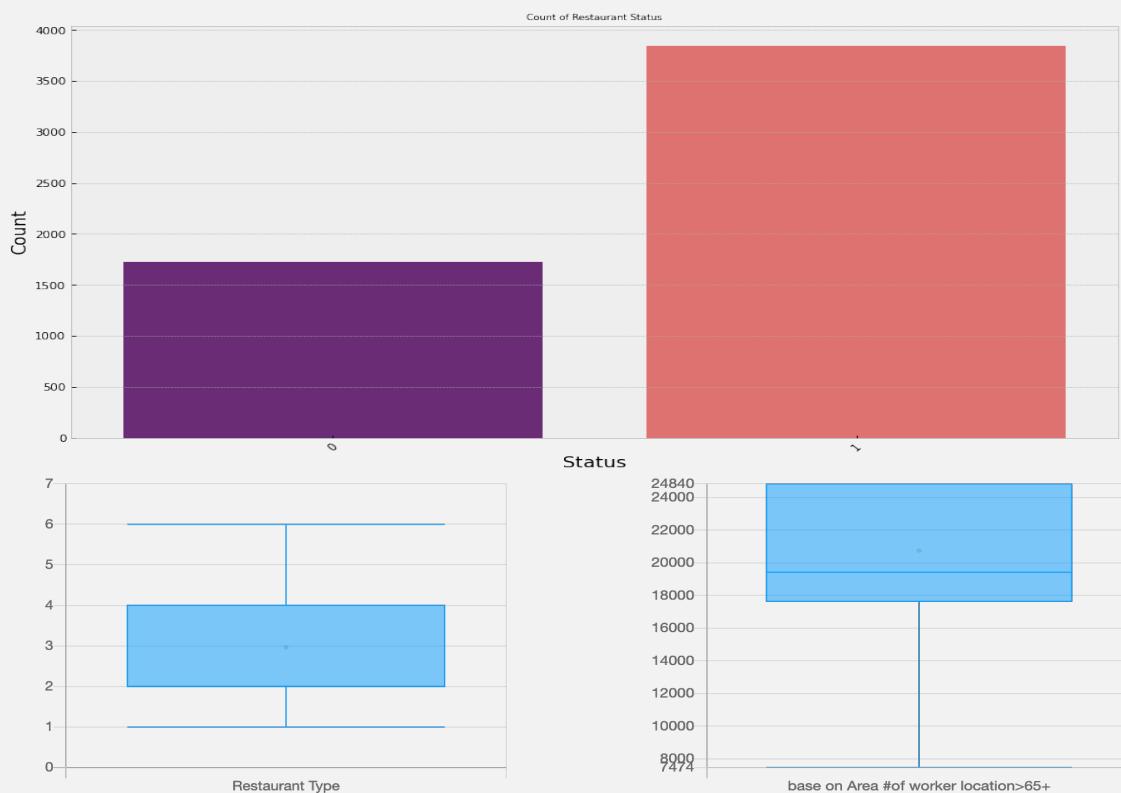
Package Management



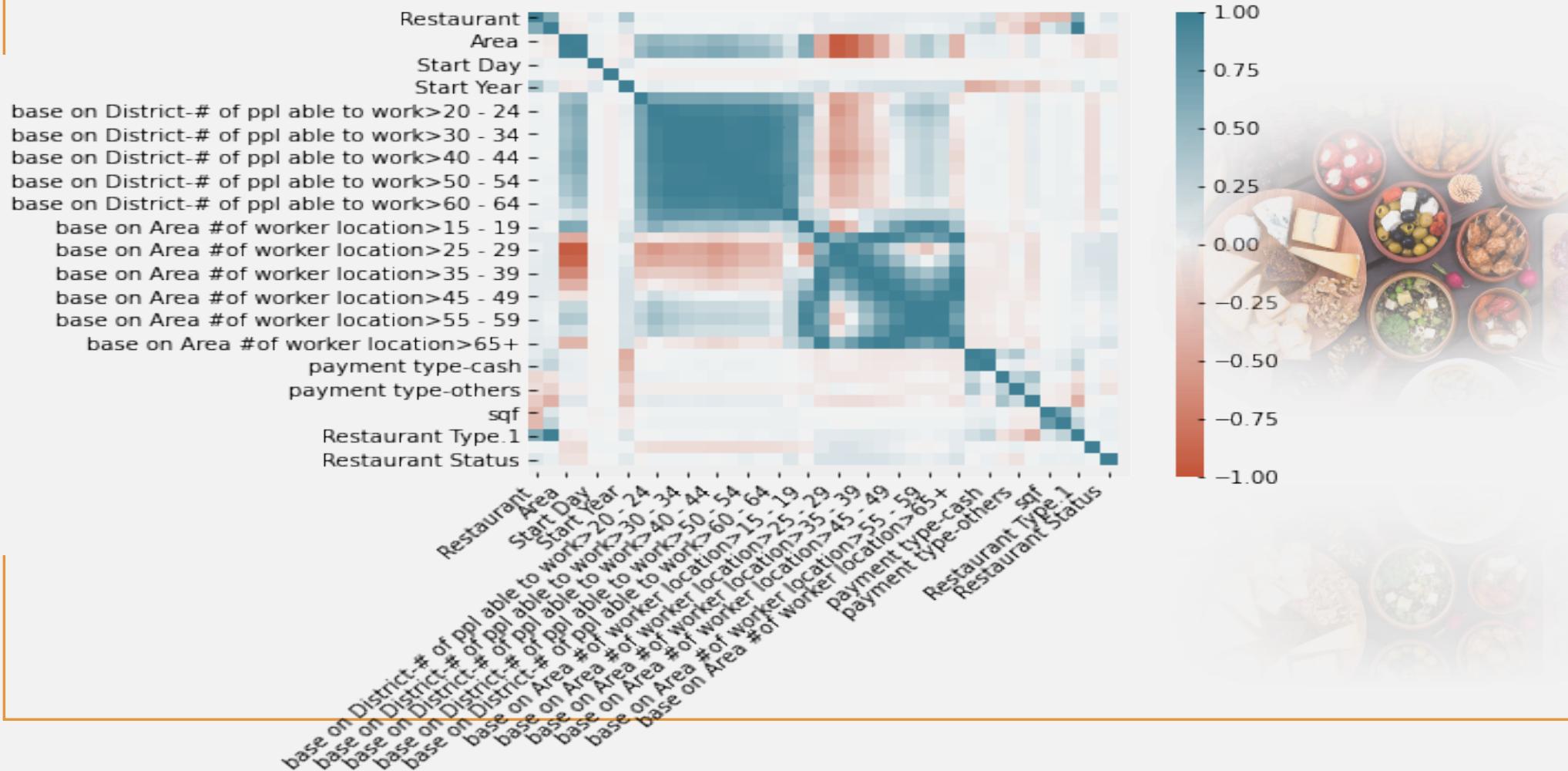
# Chapter 4 - Results and Evaluations



## Visualization of data, Analysis, Apply ML models and Evaluations



# Correlation with all features

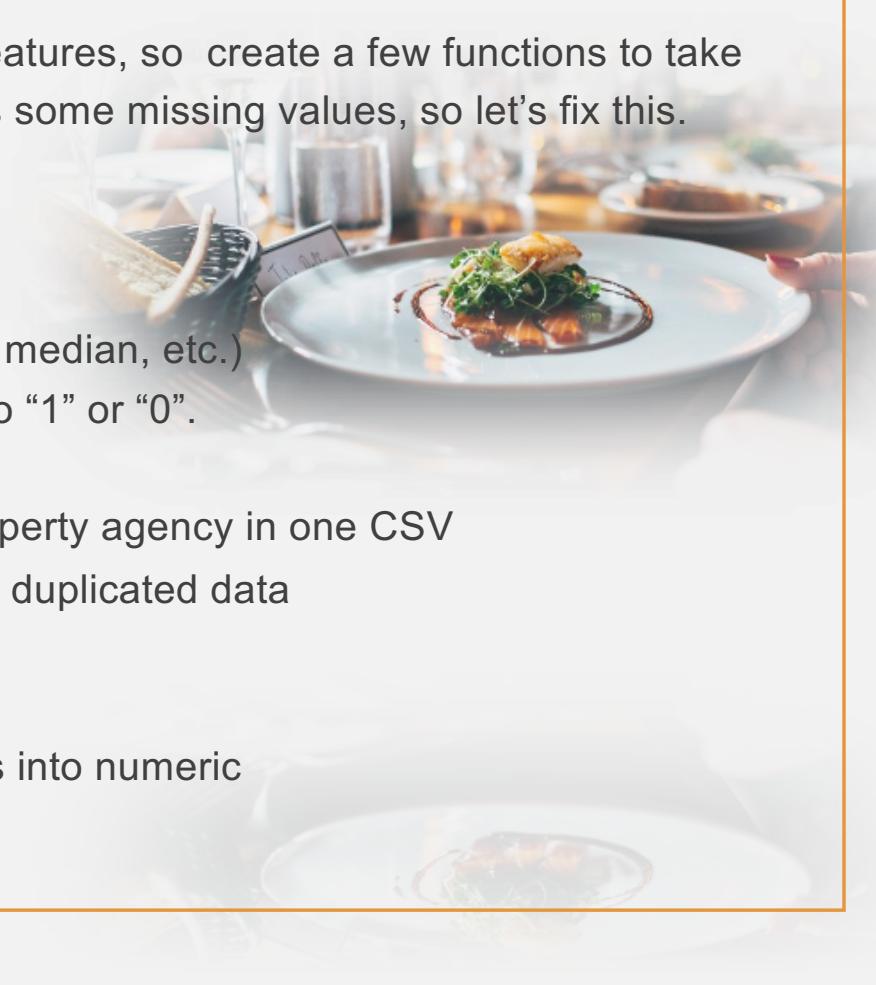


# Data Cleaning and Formatting

Machine Learning algorithms cannot work with missing features, so create a few functions to take care of them. We noticed earlier that the Sqf attribute has some missing values, so let's fix this.

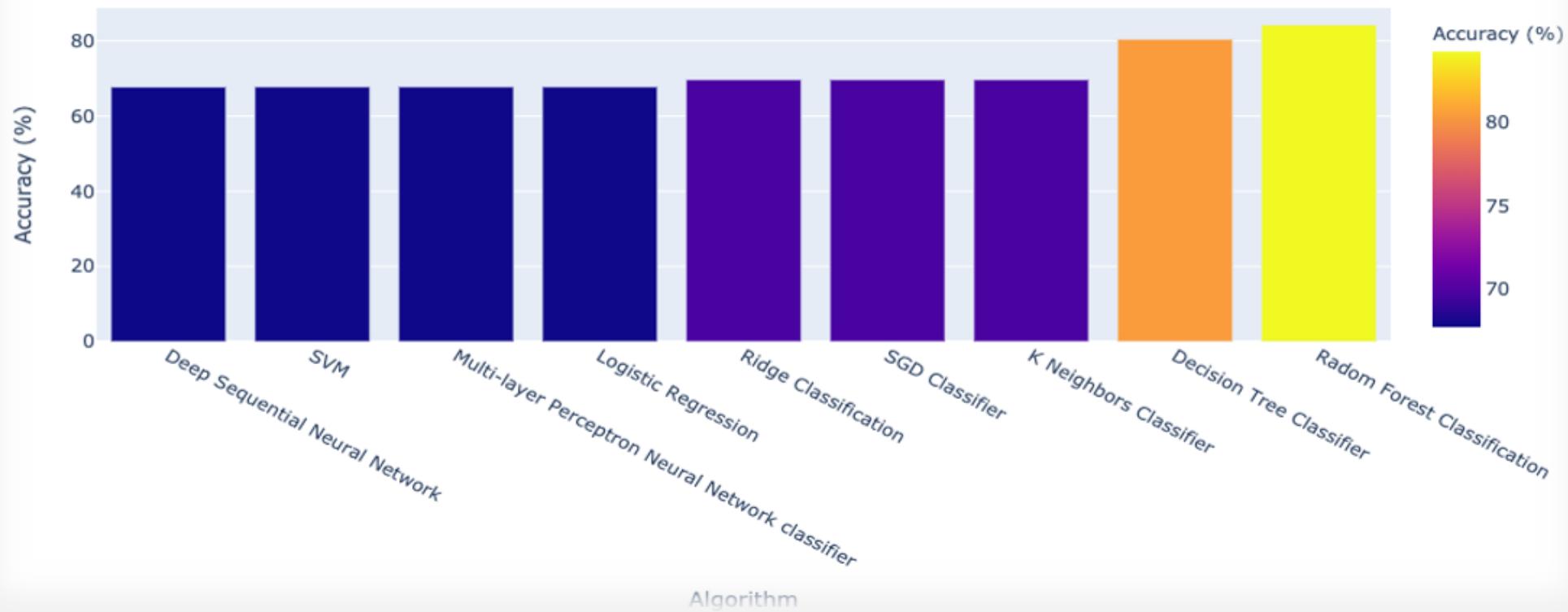
We have three options:

- Get rid of the corresponding districts.
- Get rid of the whole attribute.
- Set the Sqf values to some value (zero, the mean, the median, etc.)
- Change the restaurant status from “Open” or “Close” to “1” or “0”.
- Reformat Open and Close date into three columns
- Consolidate 3 data sets from Ambros, GovHK and Property agency in one CSV
- Some restaurants are located at same district, remove duplicated data
- Remove some attribute according to references
- Rename headings of attributes
- Convert categorical and predictor's input data columns into numeric
- Remove outliers

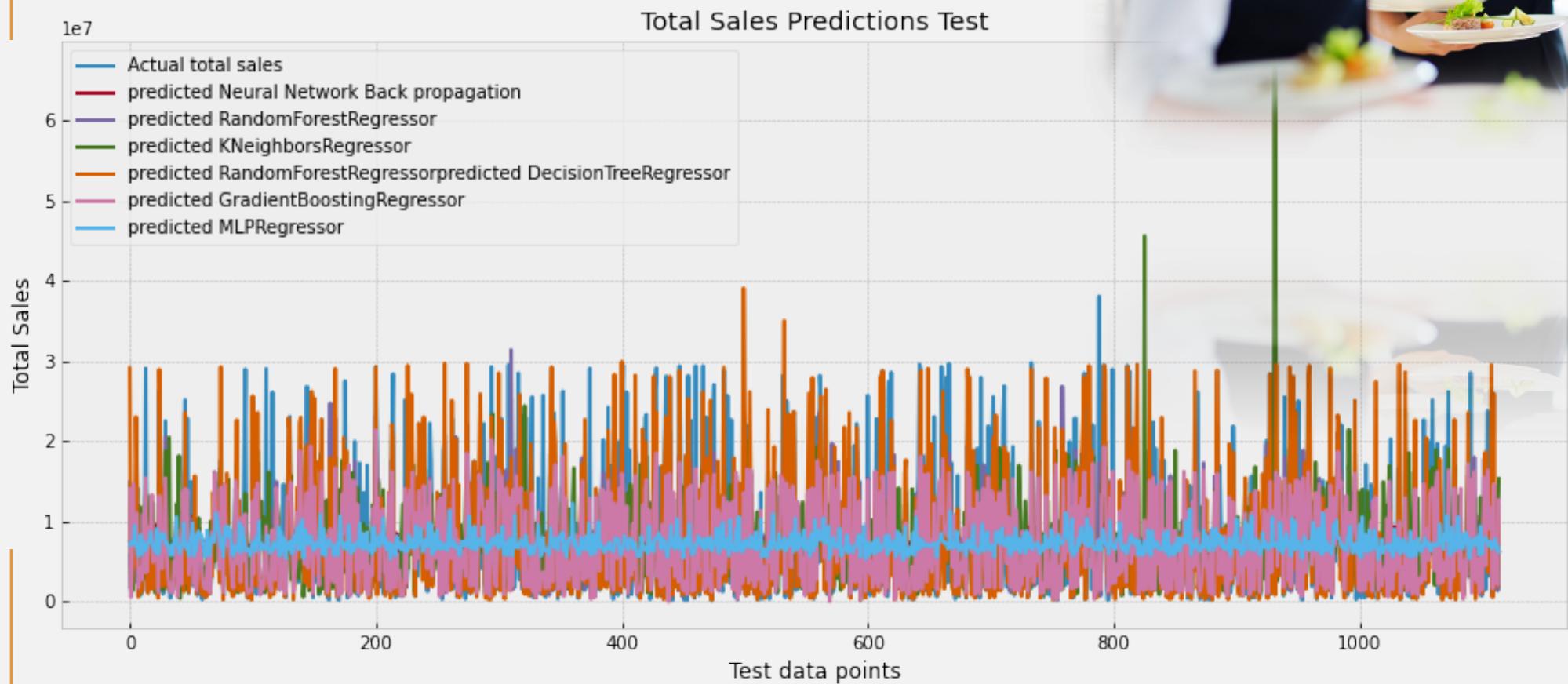


# Classifier Forecast Accuracy

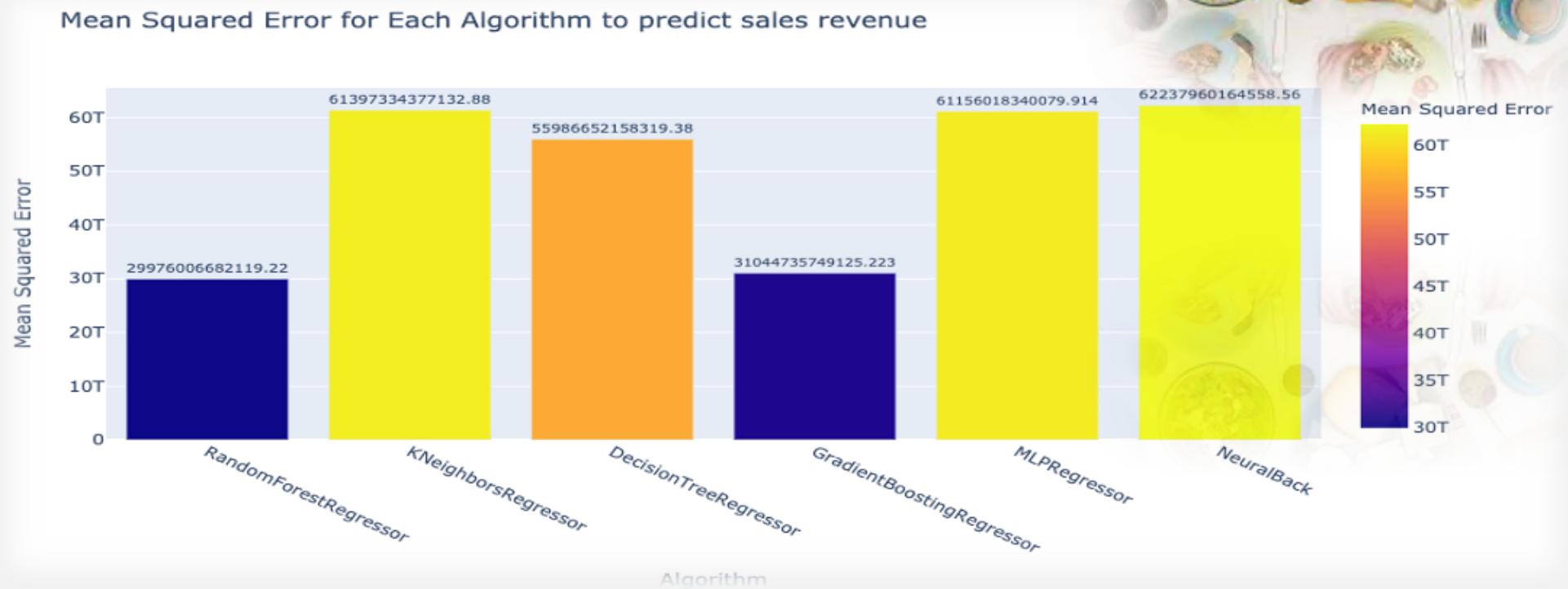
Accuracy of each Classifier for Restaurant Status Recommendation



# Total Sales Predictions Test



# Mean Squared Error for Each Algorithm to predict total sales



# Chapter 5 - Conclusions



The research has fulfilled the requirements of the client. The aim of the research is to use supervised classifications to predict restaurant's status and to use regression models to predict restaurant's sales. The literature review and complete data product development methodology were provided by this research. There are also results evaluations and visualization added. The research shows that the Random Forest Classification is the best method for restaurant's status prediction and the K Neighbour Regression is the best method for restaurant's sales prediction.

The research found that the data selection must be randomly selected for better accuracy. Random Forest Classification is the best restaurant status prediction. K Neighbour Regression is one of the best total sales predictions. Most people available for work in Hong Kong are between the ages of 20 and 40

**End**

【Machine Learning】 【Prediction】 【Algorithms】

May 2021

