

Sr#	Project Name	Description
1	Next Word Prediction-Language Model	<p>Language modeling involves predicting the next word in a sequence given the sequence of words already present. A language model is a key element in many natural language processing models such resolving customers inquiries through chat or answering the questions through emails.</p> <p>In customer Service business especially in messaging or Chats or email supports, customer representative often struggle to response fast if they have limited knowledge of business area wherein the inquiry is about and need respond fast for better service and improved customer satisfaction.</p> <p>The model that would always reads previous text and understand the context and predict the next words as agent start responding, this way they do not need to worry about appropriate language plus context of the chat. To build this model, RNN's LSTM method is used, which has a hidden state and a memory cell with three gates that are forgotten, read, and input gate. The model will be a sequential model. We are going to create embedding layer and specify the input dimensions and output dimensions, then we will add an LSTM layer to our architecture.</p> <p>The model is predicting the next words based on given input sample text, the generate next word or series of words can be handled through argument as configuration and changed based on model performance.</p> <p>The next word prediction language model is generating the relevant word and can be used in production to help agent type response fast.</p>

2

News Categorization

From the beginning, since first newspaper printed, every news that makes into page has had a specific section allotted to it. The newspaper style, news sections, format etc. have been changed over the time but not the categorization of the news and it still carried over even in to the digital version of newspaper. News articles are not limited to few topics, it covers a wide range of interest from politics to sports to movies and so on. For long time, this process of categorization news was done manually by people and used to allot news to respective section(category). With digitalization of newspaper, the news gets updated every moment and allocating to them to appropriate category can be cumbersome task.

How to Solve this problem - To avoid manual news categorization, with help of latest technology, Natural Language Processing and machine learning, this problem will tackled to classify and predict which category a piece of news will fall into based on the news headline and short description.

What model will be built for and How it would help - In order to solve the manual news categorization problem, A machine learning model will be built using supervised machine learning techniques, that would learn from existing news headlines and short description and predict the news category appropriately. With the help of this model the news categorization can be automated, and it would save manual work and help users to read the news of their interest in right section.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

It is more apt for multi-class classification task. By default,XGBClassifier or many Classifier uses objective as binary but what it does internally is classifying (one vs rest).

Walmart Sales Forecasting

Sales Forecasting is the process of using a company's sales records over the past few years to predict the short-term or long-term sales performance of the company in the future. This is one of the pillars of proper financial planning. Sales forecasting is a globally conducted corporate practice where number of objectives are identified, action-plans are chalked out as well as budgets and resources are allotted to them. Here in this project, we are going to build the Sales Forecast Model that would learn from the past sales records, events and predict the accurate sales so company will be ready to source appropriate resources before the actual event happens. The Sales Forecast Model will be machine learning model built using python, trained, and tested on Walmart sales data, the CRSP-DM methodology is used to complete this project. Sales Forecast Model is trained using different Machine Learning algorithms such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting (XGBoost) and model with highest accuracy score is baselined to predict the sales forecast in real time. The factors considered to build the model and findings about the data with the detailed use cases and technical information is provided in the paper. Walmart, Inc. is part of the retail and wholesale business and is based in Bentonville, Arkansas. The President, Chief Executive Officer, and Director is C. Douglas McMillon. Walmart operates Walmart, Walmart Neighborhood Market, Wal-Mart, Walmart.com, and Sam's Club. Retail companies commonly have issues with predicting sales accurately throughout the days, months, and years ahead. Business Sales Executives often find themselves scrambling for answers when it comes to sales forecasting during business reviews with their leaderships team. The Sales Forecast Model will help sales executives to find such answers upfront and be ready with numbers and predictions to share with leaderships team. This model would help individual stores to upscale their customer satisfaction by stocking the right products at right time and decrease overstocking and wastage of food products. The baselined model used **ML algorithm XGBoost Gradient** can be deployed to production system and simple application can be built to predict sales. The Application would accept values such store number, department number, week of the year, size of the store, is holiday in the week, average temperature, unemployment rate in that week, fuel price etc. and based on these values it would predict the sales value for that store and departments.

4

Predict House Prices in Austin TX USA

House Prices in USA are booming, and house prices will continue to race ahead, at nearly twice the pace predicted before this year. This is what we hear or read when talk about housing market in USA. Buying house is very critical job, one should be aware of lots of things before buying house and when buying house nobody sure about when is the right time to buy house and wants to have some tool that would consider all the factors determining house price and predict the house price. Predicting the house price is challenging but doable and with help of machine learning algorithms this can be achieved. This is the topic for final project - predicting the house price based on the house features, area features. The data sets I obtained are from houses in Austin TX. Housing market in USA continue to rise in all the country's major cities. The Austin TX market is also one of the hottest markets in 2021. The data set has total 47 unique features and it was uploaded with house images for each house listed in the data set and given the reference of each house to the data set as an additional feature. In order to predict the house prices from given house features and area features, the main business problems are: 1. Identify the features that impact house Price. 2. Predict the house prices based on identified features. Overall, by performing the exploratory data analysis and regression analysis, we were able to handle the problem statement mentioned above. With the help of exploratory data analysis, the first problem statement was resolved to identify the features that are impacting the overall house prices. Regression analysis helped to tackle second question wherein we need to predict the house price based on given features of house and property area. The **Multiple Linear Regression model** built to tackle second question on predicting house price and based on the result summary the model we built is statistically significant and predict house price.

5

Housing Price Prediction - Data Prep and Exploratory Data Analysis

The topic for project **Housing Price Prediction** is chosen for the final project to make data ready by collecting from different sources and cleaning, transforming, and merging all these data to make it final dataset in the ready format for machine learning algorithms so predicting house price model can be developed by training and validating on housing market dataset. To build predicting house price model, we should consider different factors such as house information, and facilities available in neighborhood such school, hospitals etc. As part of this exercise, Austin, Texas housing market data will be collected from 3 difference sources such as csv file with house related information, json file for hospital in neighborhood and tabular data for school information from website.

1. CSV File – The csv file is acquired from Kaggle website, this data set have Austin TX area house information, such as address, year built, sale price, No. of bedroom, bathrooms, lot area etc. Please refer data dictionary for more information about data. Link to Data Set -

<https://www.kaggle.com/ericpierce/austinhousingprices>

2. Tabular data – The neighborhood school information will be pulled from below websites, since the website has school information such as address, type of school, student per teacher ratio, ratings and percentile, star awards ratings etc. but we don't need grade level rating so we are going to remove them and clean data and at last will be joined with main data set based in zip code value. Links to data –

<https://www.schooldigger.com/go/TX/city/Austin/search.aspx>

3. Json File – The neighborhood public health locations information is pulled from data.gov using API. The data is in json format, and it will have public health locations information such as facility name, address, hours of operation, website etc. Data dictionary has all the detailed information about this dataset. Link to data set – <https://catalog.data.gov/dataset/austin-public-health-locations>

As part of this project data collected from different sources and performed data preparation operations and exploratory data analysis by plotting different plots to visualize the data understand the variables required for predicting house prices. This data set will be used to create model to predict the house prices, so different factors will be considered to see if those have any direct or indirect impact on house prices. All different factors will be plotted against the house price to see if there is any correlation between them and also all other features will be tested for

		multi-collinearity to make sure there are no correlation between features themselves.
--	--	---

6

**Predict
Customers
Churn in
Credit Card
Business -
Exploratory
Data Analysis**

The objective of this project was to perform the different statistical techniques as part of exploratory data analysis using python as programming language. The author of textbook 'Think Stats' has explained different techniques with real world examples. This document provides the highlight on exploratory data analysis steps performed, outcome of the project, challenges faced, assumptions made and any data missing etc. on completing the final project for this course.

Statistical/Hypothetical Question – The dataset consists of credit card customers information along with customers who closed their card and who did not. The main business problem was to predict customers who are going to churn, so that appropriate offers are given to such customers, or their issues will be handled if any. The objective of this data analysis project is to better understand the existing customers information based on different data points and identify the attributes that are making impact on customer's decision to leave the credit card company and after finalizing the attributes another goal is to predict such customers who are going to leave. The customers information such as age, income level, credit card utilization, how often they contact customer service, education etc are driving them to leave or any other factors. This project were tackled using exploratory data analysis and regression analysis. As part of exploratory data analysis, different statistical techniques and are used to explore the data and different charts are used for visualization. The data validated for any missing or duplicate values, the data distribution is checked for any outliers and removed them appropriately. Different distributions are checked for few key variables such as PMF, CDF, analytical distribution such as logarithmic distribution, checked for the correlation, covariance between the variables, plotted scatter charts to see correlation visually. This exploratory analysis helped to identify features that are impacting customers decision to leave or stay with bank. Finally, Regression analysis was performed using the key features to build the model to predict the customers who will drop-off and who will not. This exploratory analysis helped understand the different features, their relationship and impact and regression analysis helped to build prediction model more accurately. Overall, with this analysis we were able to solve the business problem that was stated to identify features impacting customer to leave or stay and predicting churn.

7	eCommerce Future - Google Trend - Research	<p>To begin with the Term project for this course and part of M1, I have decided to research topic e-Commerce. This is very broad term to research so scope of this project is to research the question - What is the future of e-Commerce? Will it continue to grow in future?</p> <p>To dig into this topic, we are going to implement the data science methodologies we learned during this course.</p> <p>First thing first, Let's implement the CRISP-DM techniques to analyze this problem. As we learned in this technique the first and very importing stage to understand the problem by gaining the knowledge about it is Business Understanding -</p> <p>What is e-Commerce? e-Commerce (or electronic commerce) is the buying and selling of goods (or services) on the internet.</p> <p>This topic of the project was solved by performing online research and using Google trend by searching keywords and comparing the trend results.</p>
8	Building Weather App	<p>For this project, it was Python coding practice to build weather application using Python by reading data from weather API. In this project built the application, that would read weather API by providing valid keys to read data based on user provided details such city name or zip code or both to fetch relevant weather information to display for user.</p>
9	Project 2 From DSC 680	
10	Project 3 From DSC 680	