KODI PRAKASH SENAPATI
LEAD DATA-SCIENTIST
MICROSOFT CERTIFIED AI ENGINEER
E-mail: kodidatascientist@gmail.com
Mob# +91 – 8309209079

PROFESSIONAL BACKGROUND

Having 10+ years of diversified IT experience in analysis, design and development of Data science (Machine Learning, Deep Learning, Artificial Intelligence, Natural Language processing, Neural Network, Computer Vision), Big Data applications (HDFS, Map-Reduce, Hive and Spark) and TABLEAU

DATA SCIENCE EXPERIENCE

- Understanding Business Problem Using Python and Bigdata concepts to solve business Problems and successful implementations of solutions.
- Possess expertise in Data science (Machine Learning & Deep Learning) & NLP expertise using Python.
- Possess expertise in implementation of Regression, classification and clustering algorithm like (Linear Regression, Multiple Liner Regression, Logistic Regression, KNN, Naive Bayes classifier, Decision Tree, Random Forest, SVM, K-Means, Hierarchical clustering, Gradient descent, ANN, CNN, RNN, Computer vision, Natural language processing and Recommendation systems, LSTM, Time series analysis) using scikit learn and tensor flow libraries.
- Possess expertise in implementation of chatbot using Google Dialog Flow and strong knowledge in NLP using Spacy.
- Good in statistical and mathematical knowledge
- Strong knowledge of Banking, Retail, and healthcare domain.
- Model development such as predication, Image recognition.

EMPLOYMENT HISTORY

- Working as Full time LEAD DATASCIENTIST at Synergistic IT since Sep 2021
- Working as part time DATASCIENCE Trainer at UPGRAD [IIT-BANGALORE] since Feb 2020
- Worked as ARTIFICIAL INTELIGENCE SOLUTION LEAD at KIABI since June 2020 to Aug 2021
- Worked as TECHNICAL LEAD DATA SCIENTIST at VALUELABS PVT LTD since June 2019 till Jan 2020.
- Worked as PROCESS DEVELOPER at GENPACT PVT LTD since AUG 2011 till FEB 2019.

EDUCATIONAL QUALIFICATION

- M.C.A(master’s in computer application) from AVIT College, CHENNAI, DEEMED University, CHENNAI–2010

TECHNICAL SKILLS

Data Visualization || Tableau || Predictive Analysis || Statistical Modeling || Classification || Clustering || Data Analytics || Data Mining || ML Algorithms || Model Development || NLP || Text Pre-processing || Tokenization || Lemmatization || Stemming || NER || Text Summarization || Topic Modeling || Sentiment Analysis || Word2Vec || Tfidf || SKlearn || TensorFlow || NLTK || SPACY || Big data || Hive || Hadoop || Deep Learning || Neural Network || mask RNN || YOLO

PROJECT DETAILS

1.	**Projects**	**:**	**Delay Order Prediction**
	Client	**:**	**KIABI**
	Duration	**:**	**June 2020 to Aug 2021**
	Role	**:**	**AI Solution Lead**

DESCRIPTION:

- KIABI is one of the French based retail industries which is mainly deals with Textile & Garments. Working for delay order prediction project, the main object is to predict the future PO & implement this model for automation to reduce the manual work.

- USE CASE: To predict the delay in days for future Order's & To classify weather the order is delay or not and also to predict the probability of delay % for future order's & implemented all classification model which is – Logistic Regression, SGD Classifier, Decision Tree, Random Forest

Responsibility:

- Being an individual contributor pulled historical report from 2014 to 2019 from database and did not consider 2020 because of uncertainty situation
- Train the data till 2018 & test the 2019 model
- As my dataset is supervised learning & my problem statement is regression problem that's why I used regression algorithm.
- Data preprocessing steps are done – Acquire data from the database. Import the libraries, import the dataset, identify and handling missing values, encoded with categorical data, Splitting the dataset, feature scaling
- To find the constant attributes which are available in my dataset
- Once we find the attribute selection split the data into train & test
- Apply machine learning regression model like – Linear Regression, Lasso Regression, Elastic -net Regression, Decision tree regressor, KNN regressor, Gradient Boosting Regressor

2.	**Projects**	**:**	**Image Similarity Matching (POC PROJECT)**
	Client	**:**	**KIABI**
	Duration	**:**	**June 2020 to Aug 2021**
	Technology	**:**	**YOLO, BCRNN GAN, CNTK, DETECTORN, ENAS ARCHITECTURE, OPEN CV, Tesseract, TensorFlow**
	Role	**:**	**AI Solution Lead**

DESCRIPTION:

Main objective is to recognize images and classify them into categories e.g. Pant, Shirt, T-shirt, sandal, sneakers etc. Then once image is classified, we further recognize the color or color combinations it has, and any other attributes like checks, polka dots strips etc, or any prints it has like animal print, flower, tree etc. After this, the use case is that if someone uploads an image, then the application can recognize the uploaded image, and display all images which are similar to it.

USE CASE FOR PROBLME STATEMENT:

- Recognize and categorize images from dataset & Pattern matching
- To recommend similar images (User upload an image model will recommend the imaged with accuracy)

Responsibility:

- In order to Categorize Images, we will first label the dataset with all possible patterns. E.g. - (if we consider shirts then Check shirt, dotted shirt, stripped shirt, Floral pattern with color combination)
- we filter out color combination and other pattens in later stage with OpenCV it will more time even though we use spark distributed framework.
- I have planned to use 4 models i.e. (BCRNN, YOLO, Mask-RNN, Detectron2) to check the accuracy with our dataset and choose best one
- In BCRNN we will convert the dataset to TensorFlow object record to compress image size and process the model.
- In yolo we will create labels and anchor calculation to increase the localization and predict the labels.
- In mask-Rnn and detectron2 will use same labelling and predict the output.
- For hyperparameter tuning and to increase accuracy if required we can ensemble 2 or 3 models
- If apart of labels any other text present over images we will use TESSARECT to extract images and if required we will use any other available models to extract images in case of less accuracy if images contain noises i.e. (blurred text present over images etc.)
- Once I got the image classification as text then i will match the text with labels and show all the image present in the label folder (i.e., recommend all the similar images)

FUTURE R&D STEPS:

- We have planned to create ENAS (Efficient Neural Architecture Search) network which will decide in runtime depending on the image which model out of 4 model will give more accuracy so we can deploy the same product across all region so that manual intervention will be reduced.

- In future planned to implement streaming so that once data available it will automatically train and test the result. If any user upload new dataset then model trained automatically & provide the results

3.	Projects	:	Mediquality Article Clustering
	Client	:	WebMD - Internet Brand
	Organization	:	Value Labs
	Duration	:	June 2019 – Jan 2020
	Technology	:	Python, Spyder, Sklearn, NLP, NLTK, spacy, xml. tree, Beautiful-soup
	Role	:	Technical Lead

DESCRIPTION

Mediquality Article clustering project is all about list of XML articles unstructured data. Worked this project using NLP packages to find the import tags. Used below steps to complete this project –

- Using Beautiful soup package slicing each XML documents to convert text document.
- Worked on data cleansing part by removed all Stop words, Non-Ascii character, de-noise characters.
- After data cleansing convert entire text document to list of single words.
- Extracting important keywords from text documents after done with lemmatized and stemming words.
- Word cloud is a data visualization technique used for representing text data in which the size of each words indicates its frequency or importance, significant textual data points can be highlighted using word cloud.
- Word clouds are widely used for analyzing data from social network websites or articles.
- After keyword extraction worked with NER (Named Entity Recognition) to recognize named entities using classifier and the classifier adds category labels such as person, organization and Location.
- Based the client requirement I have to find the sentence scoring and using text summarization I completed sentence scoring with ranking order and sentence scoring 1st step is – converted entire document to list of sentences, 2nd step – Text preprocessing and tokenize the sentences 3rd step – Find Weighted frequency of occurrence to get token scoring, 3rd steps – Replace words by weighted frequency in original sentences to get sentence scoring.
- Convert entire text document to sentences to find the sentence scoring, based on scoring of each sentence we can figure out which sentence have high importance of a document.
- Cluster (Grouping) these keywords together – Applied K-MEANS clustering algorithm to build and fit the model to cluster the unsupervised XML data.
- Applied ELBOW method to find out how many groups can be created and using elbow method to figure out the K-value and the elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of cluster in a dataset. As per the graph if the line is slightly curve at which point, based on that point we can create those many groups.
- As the data is unsupervised hence, I used K-MEANS Algorithm and Cluster visualization done using word cloud.
- Also using spacy package to find the location and money. Created data frames to identify which no. of articles belongs to which group.

RESPONSIBILITIES

Done exploratory analysis and took inferences by visualization the data || Made exploratory analysis and cleansed the data || Performed below are the data preprocessing steps || Tokenization || Noise Removal || Noun Phrase Extraction || Part-of-speech Tagging || Words Inflection and Lemmatization || Stemming || Unigram || Bi-gram || N-grams || Text summarization & Topic modeling || K-MEANS Clustering algorithm || Text summarization & Topic modeling || Used packages- NLTK, SPACY, XML.ETREE, BEAUTIFULLSOUP,GLOB

4.	Projects	:	Lima chatbot
	Duration	:	Feb 2018–Feb 2019
	Organization	:	GENPACT
	Technology	:	Tensor-flow, Keras, Theano, oracle, Anaconda package, Google API, Google Dialog flow, Google Cloud
	Role	:	Team Lead

Chatterbot is a machine-learning based conversational dialog engine build in Python which makes it possible to generate responses based on collections of known conversations. The language independent design of Chatterbot allows it to be trained to speak any language

An untrained instance of Chatterbot starts off with no knowledge of how to communicate. Each time a user enters a statement, the library saves the text that they entered and the text that the statement was in response to. As Chatterbot receives more input the number of responses that it can reply and the accuracy of each response in relation to the input statement increase. The program selects the closest matching response by searching for the closest matching known statement that matches the input, it then returns the most likely response to that statement based on how frequently each response is issued by the people the bot communicates with

For building the Bot we used Google's natural language understanding developer framework for building conversational experiences. Dialog Flow needs to be trained on the dataset to attain a machine learning capability which understands the intent and context of what a user says in order to respond in the most useful way.

Dialog Flow lets you build conversational interfaces on top of your products and services by providing a powerful natural language understanding (NLU) engine to process and understand natural language input.

RESPONSIBILITIES

Done exploratory analysis and took inferences by visualization the data || Made exploratory analysis and cleansed the data || Performed below are the data preprocessing steps || Tokenization || Noise Removal || Noun Phrase Extraction || Part-of-speech Tagging || Words Inflection and Lemmatization ||
N-grams || Spelling Correction || Semantic Analysis || Sentiment Analysis || Performed below are the Training Phases using Google Dialog Flow || Content Recognition || Content Level Division || Entities
Events Generation || Synonym Tuning || Response Training || Using Google cloud (GCI) performed Bot

5.	**Projects**	**:**	**Advisory Group - Advisor Retention Strategy Support**
	Duration	**:**	**Apr 2017 – Jan 2018**
	Organization	**:**	**GENPACT**
	Role	**:**	**Module Lead**

DESCRIPTION

AIG (American international Group) is an American multinational finance and insurance corporation. The company operates through three core businesses: General Insurance, Life & Retirement, and a standalone technology-enabled subsidiary. General Insurance includes Commercial, Personal Insurance, U.S. and International field operations. Life & Retirement includes Group Retirement, Individual Retirement, Life, and Institutional Markets.

- Done with the data preprocessing and Exploratory data analysis
- Splitting data into training and testing data to consider 80% data as training data and the remaining 20% of data for testing
- Based on result outcome we may change the above ratio
- Methods to use for minimizing the Bias and Variance of the algorithm using Bagging, Boosting & Voting
- Algorithms used are – Logistic Regression, SVM, Random forest & Neural Networks
- Finally identify the major influencing factors and model can predict the terminated advisors

PROBLEM STATEMENT – Prepare the Advisors retention strategy report.

RESOLUTOIN:

- Identifying the probable advisors who are likely to leave the organization.
- Identifying the most influencing factors which will impact advisors to leave the organization.

6.	**Projects**	**:**	-CDD -LOANIQ -Loan defaulter prediction -Customer feedback analysis
	Duration	**:**	OCT 2015 – MAR 2017
	Organization	**:**	GENPACT
	Role	**:**	Process Developer

DESCRIPTION

Wells Fargo & Company is an American multinational banking and financial services holding company which is headquartered in San Francisco, California, with "hub quarters" throughout the country. It is the fourth largest bank in the U.S. by assets and the largest bank by market capitalization. Wells Fargo is the second largest bank in deposits, home mortgage servicing, and debit cards

CUSTOMER DUE DELIGENCE (CDD) PREDICTION

CDD information comprises the facts about a customer that should enable an organization to assess the extent to which the customer exposes it to a range of risks. These risks include money laundering and terrorist financing. Primary goal of CDD enables Wells Fargo to know its customer understand the nature and purpose of customer relationship to develop a customer risk profile and reasonably predict the types of transaction in which a customer is likely to engage and determine when transition is potentially suspicious. Organizations need to 'know their customers' for a number of reasons:

- to comply with the requirements of relevant legislation and regulation
- to help the firm, at the time the due diligence is carried out, to be reasonably certain that the customers are who they say they are, and that it is appropriate
- to provide them with the products or services requested
- to guard against fraud, including impersonation and identity fraud
- to help the organization to identify, during the course of a continuing relationship, what is unusual and to enable the unusual to be examined.
- if unusual events do not have a commercial or otherwise straightforward rationale, they may involve money laundering, fraud, or handling criminal or terrorist property
- to enable the organization to assist law enforcement, by providing available
- Information on customers being investigated following the making of a suspicion report to the Financial Intelligence Unit (FIU)

LOAN IQ PREDICTION

Loan-IQ allows you to instantly identify high-risk loans and reduce overall default exposure. The Loan-IQ and Market Risk Scores assist you in making a review decision based on the level of collateral risk associated with a loan. Whether you need to assess one loan or thousands, Loan-IQ delivers the answers you need fast to make the most informed loan decisions. Primary goal of Loan IQ prediction is

- Easy-to-use and designed to work with your existing loan systems and processes
- Fast-track low risk loans to automated approval processes
- Select the highest risk loans for quality control and due diligence
- Evaluate existing portfolios retroactively for loan quality comparisons
- Monitor portfolio performance to aid with loss mitigation and retention programs
- Investor protection with insurance option
- The industry's only patent-pending collateral risk predictive model
- Loan scoring based on the widest range of predictors
- Instant identification of property over-valuation
- Filter out high-risk loans by drilling down on red-flagged indicators to assess quality of loans against your underwriting criteria
- Instantly receive risk metrics for due diligence

LOAN DEFAULTER PREDICTION

The goal of this project is to build a machine learning model that can predict if a person will default on the loan based on the loan and personal information provided. The model is intended to be used as a reference tool for the client and his financial institution to help make decisions on issuing loans, so that the risk can be lowered, and the profit can be maximized & Built a model to predict whether the banking customer will repay the loan amount or not.

CUSTOMER FEEDBACK ANALYSIS

Feedback analysis involves identifying the needs and frustrations of customers, so that businesses can improve customer satisfaction and reduce churn. It's often done automatically, enabling companies to sort huge amounts of data from various channels in a timely and accurate way.

An NLP engine scrapes customers' conversations and feedback on websites, social media channels, and mobile apps, etc., and then analyze these conversations. Depending on the intelligence required, the NLP engine can conduct binary (positive/negative) analysis and sentiment analysis on various emotions like Happy, Angry, Sadness, and Trust, etc. It can also analyze on the basis of frequency of a particular word and can also conduct word grouping to identify the different terms that often occur together like 'Affordable product', costly service'