Analysis Of Farmer Queries Data

Sukesh Bukka
CSE, BITS Pilani Hyderabad Campus
Hyderabad, Telangana, India
f20170008@hyderabad.bits-pilani.ac.in

Abhijeeth B V
CSE, BITS Pilani Hyderabad Campus Hyderabad, Telangana, India
f20170026@hyderabad.bits-pilani.ac.in

Surya Sidhartha S
CSE, BITS Pilani Hyderabad Campus
Hyderabad, Telangana, India
f20170200@hyderabad.bits-pilani.ac.in

Bhaswath Narkedamilly
CSE, BITS Pilani Hyderabad Campus
Hyderabad, Telangana, India
f20170033@hyderabad.bits-pilani.ac.in

***Abstract*—Kisan Call Centers is a scheme launched by Ministry of Agriculture & Farmers Welfare in January 2004. This is to address the queries of farmers in local languages using a toll-free number. A huge amount of data is being generated daily from this. This data is used to extract useful insights using data mining techniques which can be used to further improvise the scheme by looking at the results.**

I. Introduction

In India, due to an increase in consumption demands, the agricultural sector needs to grow at a much faster rate. Most of India's workforce is dependent on agriculture directly or indirectly. The organized and systematic cultivation will increase the crop production and as a result the prosperity of the farmer. Most of the queries by farmers are answered through a service run by the government which is Kisan Call Center. The objective of the call center is to answer the needs of the farming community through this toll-free phone number, sometimes assisted by specialists in a particular agriculture field. Huge amounts of data is being generated at the Kisan Call Center, so there is a need to refine, process, store the data and find interesting patterns. These results can help the farmer to know the problems which they might face frequently.

II. Dataset Description

Kisan Call Centers are storing huge amounts of data, where the data contains the queries asked by the farmers and the responses given by Kisan Call Center operators. With many queries being asked every day, it amounted to massive data. This massive data can be used efficiently and effectively in the future if it is properly analyzed. Based on the specific regional conditions and needs of farmers, more accurate solutions can be provided to their problems with the help of this Kisan Call Center data.

The dataset contains the following columns :

Season: Information about the season. Eg: Rabi, Kharif, etc.

Sector: Information about the field in farming. E.g: Agriculture, Horticulture, Fisheries, Animal Husbandry, etc.

Category: Subcategory of Sector, specifying the type of crop the query is about.

Crop: Subcategory of Category, specifying the crop name on which the query is about.

Query type: Information about the kind of problem related to the crop.

StateName: State name.

District Name: District name in the state.

Block Name: Block in the district.

III. Preproccessing

A. *Data Cleaning:*

As the dataset contained some integer values for category types which seems invalid. So, we cleaned the data by eliminating the rows containing such values.

B. *Feature Splitting:*

From the Datetime object obtained from parsing the timestamp string we split it into two features "Month" and "Hour of the Day".

C. *Subset Selection:*

As the column Season contains almost all zero values it is considered as an irrelevant feature and is dropped. For geo-visualizations we are working towards analysing the data at the state level.

D. *Discretization:*

We discretized the data in terms of months as it would be easy for analysis and comparing the queries month wise, we named that new column as "Month".

E. *Sampling:*

Since merging all the raw datasets amounted to huge number of entries, we have decided to reduce the dataset by sampling it using "Sampling without replacement" technique.

IV. Association Rule Mining

Association rule Mining is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. In order to select interesting rules from the set of all possible rules, constraints are minimum thresholds on support and confidence, where

Support is an indication of how frequently the itemset appears in the dataset. The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

$$\text{supp}(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Confidence is an indication of how often the rule has been found to be true.

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X)$$

Some of the rules obtained using Association Rule Mining are obvious considering our dataset. Since, there are District Name and Block Name columns in our dataset.
By changing the parameters confidence and support, the **number of rules obtained were varied as shown below:**

As we can observe from the above picture obvious rules such as this district belongs to this state have a confidence of 1.

Example of obvious rules:

DistrictName_KOTA ➔ StateName_RAJASTHAN

DistrictName_ADILABAD ➔ StateName_TELANGANA

Crop_PADDY ➔ Category_CEREALS

However, some of the interesting rules we obtained are:

{StateName_TELANGANA, DistrictName_ADILABAD} ➔ QueryType_WEATHER

Category_VEGETABLES ➔ QueryType_PLANT PROTECTION

StateName_UTTAR PRADESH➔ QueryType_WEATHER

StateName_WEST BENGAL ➔ QueryType_PLANT PROTECTION

StateName_RAJASTHAN ➔ QueryType_WEATHER

From the above rules we can infer that from various states such as Rajasthan, Telangana, Uttar Pradesh the most frequent QueryType was weather as it is relevant that usually most of the farmers have weather related queries, so that they can decide upon the plant they want to cultivate that particular time. As we can see that when the category is vegetables then the Query is about Plant Protection. Since, vegetable plants are more prone to pests the queries are usually about Plant Protection, and most of these plant protection Queries are from the State West Bengal.

V. Decision Tree Classification

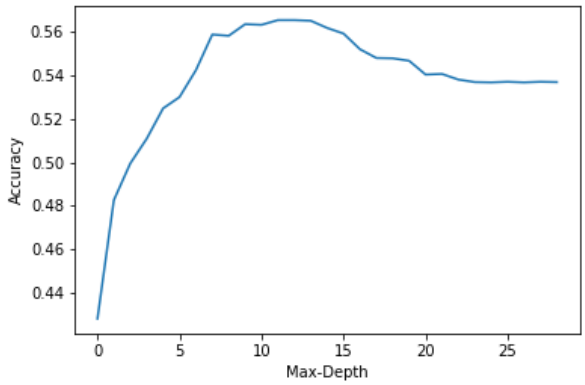
Classification is a process of categorizing a given set of data into classes, it can be performed on both structured or unstructured data. Here we used classification to predict Type of Query from Sector, Crop, Category, Location and Hour of Day attributes.

Since all feature attributes are nominal, we have used **Decision Tree Classifier** for classification. Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node.

Decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values. We use entropy to calculate the homogeneity of sample at each split. Entropy ranges from 0 to 1 completely homogeneous represented by 0 whereas if sample is equally divided entropy is 1.

$$E(T,X) = \sum_{c \in X} P(c)E(c) \qquad E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

The model is optimized based on the max-depth of tree to avoid the model being over-fitted or under-fitted. The graph shown below indicates Accuracy vs Depth from which optimal depth is obtained producing maximal accuracy.



The model generated has an accuracy of approximately 57%. This model helps in predicting QueryType which has varied uses like future query predictions and helps call centre in improving knowledge on the same for better query resolvance.

VI. VISUALIZATIONS

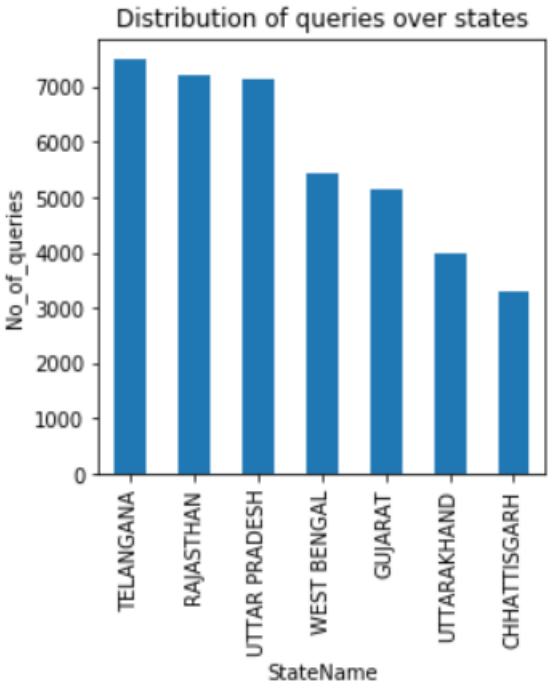


Fig. 1. Data before sampling

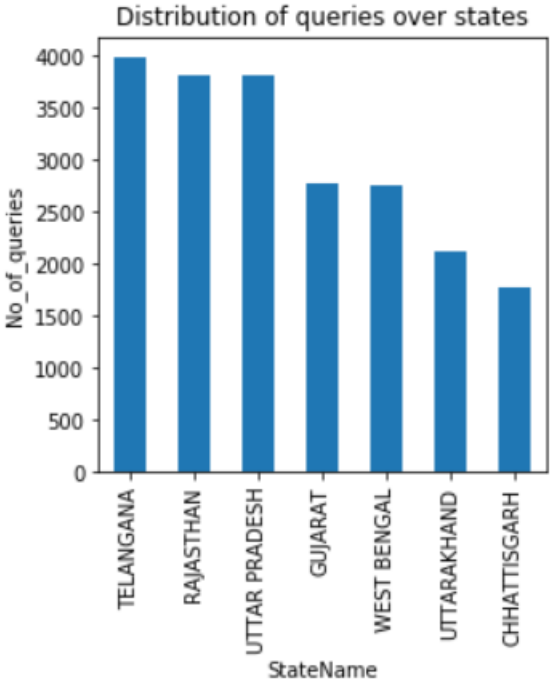


Fig. 2. Data after sampling

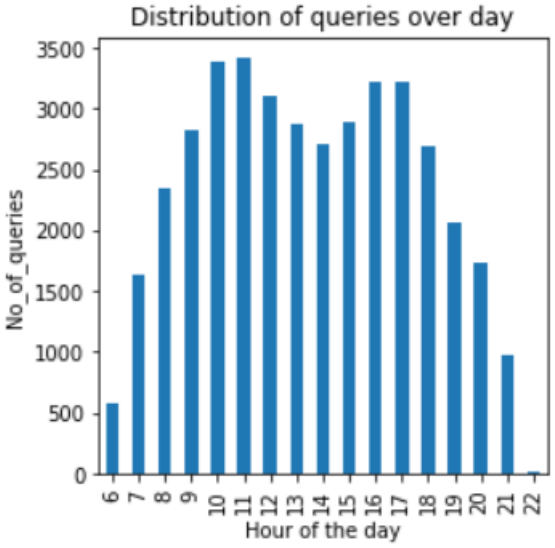


Fig. 3. Distribution of queries over the hours of the day

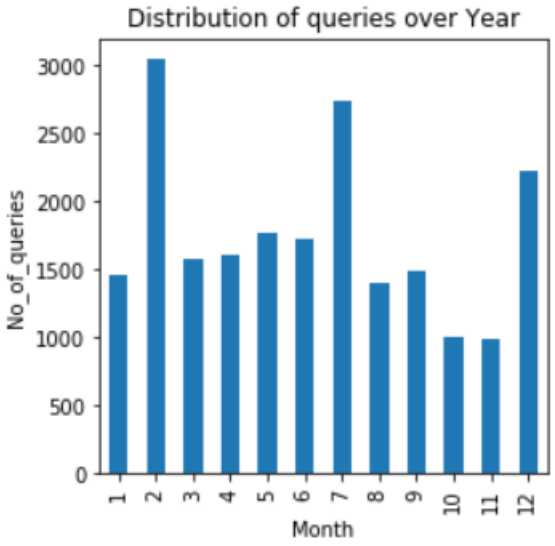


Fig. 4. Distribution of queries over the months of the year

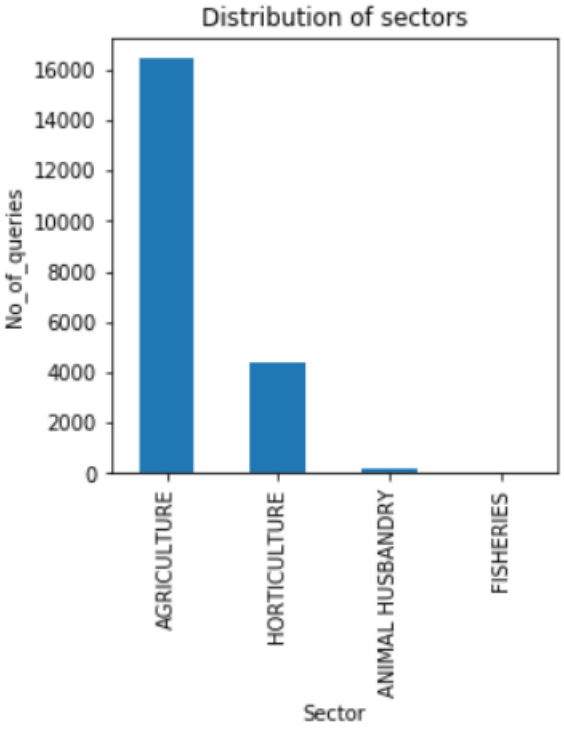


Fig. 5. Distribution of queries over various sectors

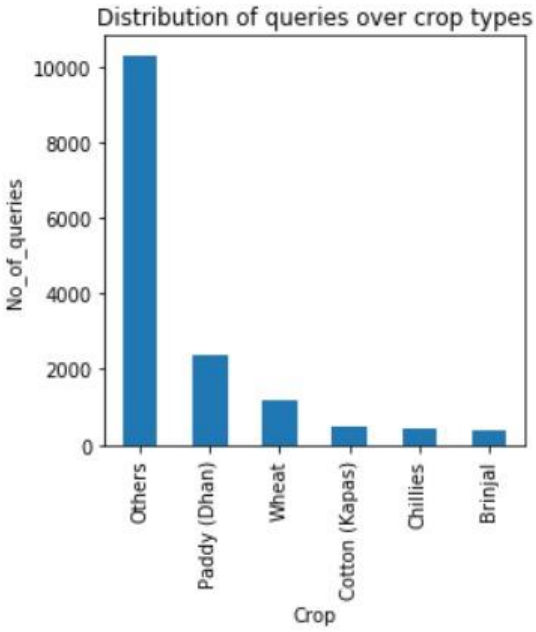


Fig. 6. Distribution of queries over various crops

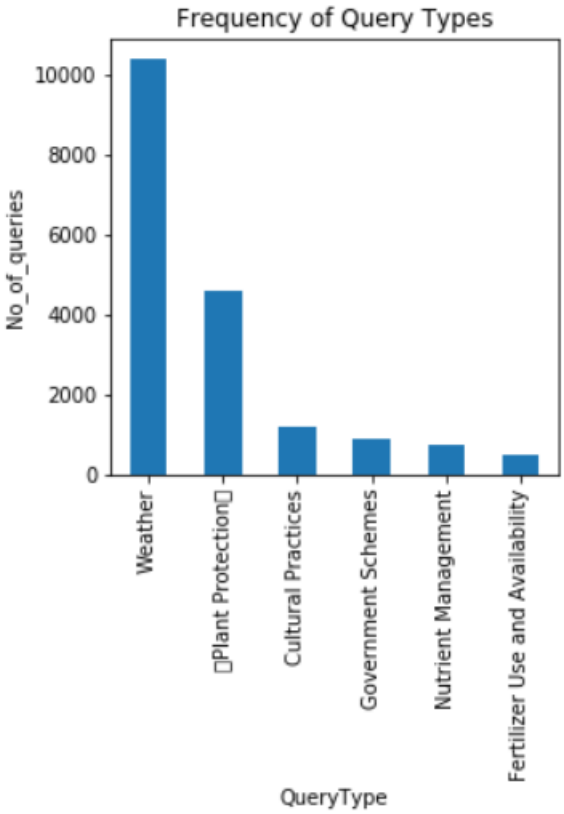


Fig. 7. Distribution of queries over various query types

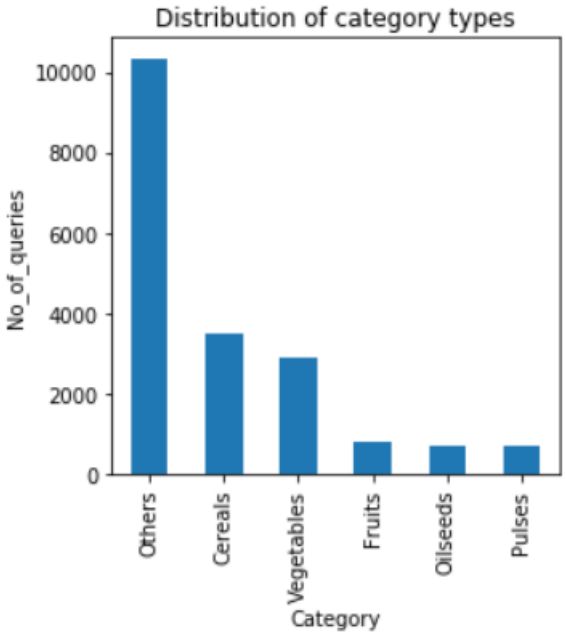


Fig. 8. Distribution od queries over various categories

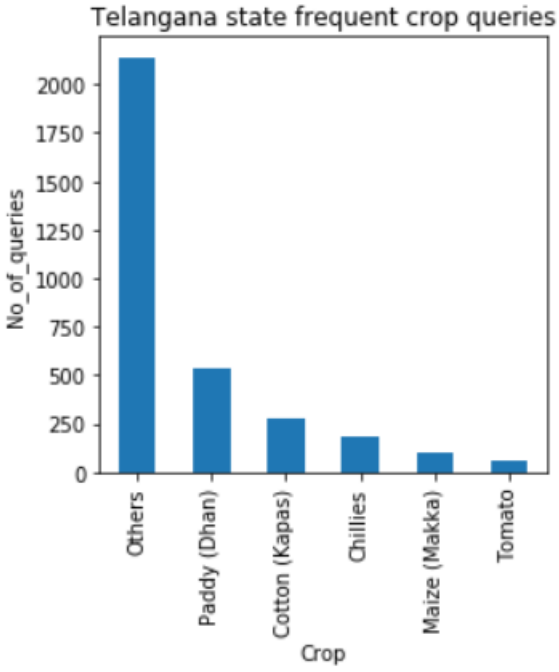


Fig. 9. Distribution of queries over various crops

VII. INFERENCES

Fig. 1. shows the pattern of queries before sampling and Fig. 2. shows the pattern after sampling. As we can see, the pattern hasn't changed considerably. That implies that our sampling worked out well and the sampled data can be now used.

Fig. 3. shows the general distribution of no. of queries during the hours of a day. As we can see, the peak hours are around 9 AM to 6 PM, and the dip in the afternoon, around 1 PM to 3 PM suggests the farmers could be having their mid-day meal. So, from this, it would be wise for the Kisan Call Centre workers to schedule their lunch break around this time, after hefty morning shift.

Fig. 4. shows the general distribution of no. of queries during the months of a year. As we can see, there is a peak at the 2nd month (Feb) which is when Rabi are harvested and Zaid Rabi crops are sown. And similarly, the peak at 7th month (July) is when Kharif and Zaid Kharif crops are sown and Rabi crops are harvested. So, it is advisable to maintain the experts related to the corresponding crop seasons at the call centers to handle the queries better, looking at this pattern.

Fig. 5. shows the top 6 sectors on which are queried about. Agriculture tops the chart, which is obvious, followed by horticulture and animal husbandry. Since the majority of queries are agriculture-related, maintaining more experts from agriculture field is appropriate.

Fig.6. shows the top crops which are queried. Since the data isn't very specific, there are a lot of items falling into others category. Among the available crops, paddy, wheat and cotton are the most queried.

Fig. 7. shows the frequencies of various type of queries. Mostly, the farmers query about the weather forecast and plant protection. So, accurate weather reports and good knowledge about plant protection techniques is expected to be maintained at the call centers.

Fig. 8. shows the frequencies of the top 6 queried crop categories. Among known categories, cereals and vegetables are at the top as they are the most commonly grown crops and the most consumed in India.

Most queried crops for each state are visualized. Fig. 9. shows the frequencies of queries on the top 6 crops in Telangana state. Paddy and cotton are the most grown crops in Telangana, which is justified by the number of queries they get. Similar visualization for other states can be found here in our GitHub repo.

VIII. References

- GitHub link to our repo
- https://en.wikipedia.org/wiki/Association_rule_learning
- https://en.wikipedia.org/wiki/Decision_tree
- https://dackkms.gov.in/account/aboutus.aspx
- https://towardsdatascience.com/market-basket-analysis-knowledge-discovery-in-database-simplistic-approach-dc41659e1558
- https://geoawesomeness.com/top-19-online-geovisualization-tools-apis-libraries-beautiful-maps/
- https://medium.com/datadriveninvestor/visualising-geospatial-data-with-python-d3b1c519f31
- https://raw.githubusercontent.com/sarangjaiswal/visualize-indian-map/master/data/state.json
- https://gist.github.com/ProProgrammer/781d5fbcb1d4364616c5?short_path=ac36090