PPT

Hello, my name is Aditya Kumar Singh. I am here to present my Data Mining and Machine learning project for CA.

The aim of the project to investigate and compare the performance of different machine learning (ML) methods as applied to three different data sets. The datasets originate from Google Projects and services. Google is an American multinational technology company that specializes in Internet-related services and products, which include online advertising technologies, a search engine, cloud computing, software, and hardware.

The Project sunroof is an initiative to calculate solar energy potential to via aerial imagery. Regression model was applied to predict solar potential for all roof types in a region. The initiative employs aerial and satellite photography to map a region and provide potential solar panel location on roofs and land. The data from Google Maps and Google Earth is used to calculate shadows and dark areas from nearby infrastructure and landscapes to find the optimal location to set up solar panels.

The project sunroof dataset was chosen because of the large number of rows and columns as well-organized records.

The motivation is the importance to predict potential benefits before any investment is placed. Hence considering the immense initial costs setting up solar panels means that enough study needs to be conducted before any infrastructure is placed. Predicting the total energy potential that can be captured is of utmost important for a feasible project.

Play store is the certified application store for android certified mobile devices Google android operating system accounted for87% of all certified mobile globally. Google Play store being the official pre-installed application store for android is the most used android app in the world. It also holds 2.65 Million downloadable application. The google dataset was chosen because of the immense records and information.

User Ratings is the biggest feedback factor which is used by developers to evaluate their applications. Hence predicting the User rating is very important in terms of ad revenue which is usually the main source of income in app development. Successfully predicting the rating will also give insight on the most co-related parameter that makes an app successful.

YouTube is online video platform with millions of stream able videos. Headquartered in San Bruno, YouTube is the largest video streaming service in the world. With millions of videos available online and millions of people watching regularly,

YouTube has been used small and large production companies to grow audiences. For this project, T-Series channel dataset was used to implement ML model. T-Series is a Music Production Company in India with the highest subscribers count on YouTube globally.

For channels on YouTube, the source of revenue stream from the number of audiences who actively watch their videos. To find the most successful genres of video it is important to classify the best category of videos.

The dataset was checked for missing or null values. All rows that showed otherwise were dropped.

Irrelevant and unnecessary features were dropped from data. columns were trans-formed into numeric values using label encoder. Correlation Matrix graph was plotted to find the correlated rows with the target variable.

The columns that showed collinearity of 1 with target variable were dropped from the datasets as it would not be impossible to implement a model. Several other columns which showed multicollinearity were dropped for as to not obtain an. Columns that showed multicollinearity with each other were also found and dropped.

The data was divided into test and train for models to fit. For this new data frames X and y were created where X contains predictors and y containers target variables. It is necessary to implement this to get parameters for each model.

Here are the combined results of Models. We can compare the performances with certain metrics.

F-score is the [geometric mean](https://deepai.org/machine-learning-glossary-and-terms/geometric-mean) of precision and recall.  So higher the F score better the model. With this we can tell