G10 Project Proposal

> Luwei Wang (lw3511) / Pallabi Chandra (pc3131) /
>
> Matias Gutierrez (meg10001) / Sunny Shah (sbs8673)

Movie Recommendation System

Business analysis

Problem description

In this world of ever-changing consumer trends, it is extremely important to understand their preferences and the way they change over time so that businesses can make recommendations to consumers regarding products/items they would like, so as to improve customer satisfaction as well as improve chances of sales and subscriptions. Thus, several businesses are investing in building their own recommendation engine. One such domain where such a recommendation system is very important is an over-the-top (OTT) streaming platform (for example Netflix, HBO Max, etc.) to suggest new movie/series titles to users based on the content they consume for customer satisfaction so that they retain their subscription and continue to pay the fee for the company's monetary benefit.

Expected results and actions

We aim to build a recommendation system based on user ratings of different movies as given by a number of different users. Based on this model, we can then make recommendations to users based on their interests once they rate films on our streaming service. Once we get these recommendations, the streaming service can then suggest titles to the consumers and verify whether the customer consumed it and gave it a high rating (the simplest real-world success case scenario for a recommendation system).

Data science problem

Prediction target

This data science problem is a supervised learning problem. We have labeled data with ratings for movie titles from specific users, along with features of the movies. The target variable for our recommendation system would be ratings for movies that have not been consumed by the user. We would use this to rank the recommendations to the user and pick top five movie titles as the output of the model.

Dataset Description

We pick three datasets to operate the model.

- MovieTweetings Dataset: This is a real-world dataset that contains ratings collected from tweets where users have mentioned their ratings for a particular movie on IMDB. This dataset consists of the following three tables:
  - Items.dat: *movie_id, movie_title, movie_year, movie genre*
  - Ratings.dat: *user_id, movie_id, rating, rating_timestamp*
  - Users.dat: *user_id, twitter_id

  https://github.com/sidooms/MovieTweetings

- MovieLens Dataset: This is a widely used dataset in the field of recommender systems, containing movie ratings and metadata. This dataset consists of the following two tables:
    - movies.csv: *movieId, title, genres*
    - ratings.csv: *userId, movieId, rating, timestamp*

  https://www.kaggle.com/datasets/parasharmanas/movie-recommendation-system

Conclusion

This proposal aims to solve a real-world data science problem. It is based on dependable datasets and has already been widely studied, which provides us with rich resources to learn the algorithms. The modeling work could be complex, but it also enables us to obtain more experience from this project.

However, the recommendation algorithms have been researched for over a decade until they achieved today's results. As a team of beginners, the prediction result may have low accuracy if we work on this project.

Diabetes Risk Prediction

Business analysis

Problem description

The business problem at hand is to assist insurance companies in determining the risk levels of potential customers, to accurately price their premiums on insurance products. This involves predicting whether an individual is at high risk of developing diabetes based on their health attributes.

Expected results and actions

Through this analysis, we aim to gain insights into the relationship between various health attributes and the likelihood of developing diabetes. Specifically, we expect to identify significant factors that contribute to diabetes risk, understand their relative importance, and explore potential interactions among these factors. Additionally, we aim to develop a predictive model capable of accurately classifying individuals into high or low risk categories based on their health profiles.

The results of this analysis will empower insurance companies to make data-driven decisions regarding premium pricing and risk assessment. By accurately identifying individuals at high risk of developing diabetes, insurers can offer targeted interventions such as wellness programs or personalized healthcare plans to mitigate risks and promote healthier lifestyles. Moreover, precise risk assessment enables insurers to optimize their pricing strategies, ensuring fair premiums for both customers and the company.

Data science problem

Prediction target

This project constitutes a classification problem in the domain of supervised learning. We aim to classify individuals into two categories: high risk and low risk of developing diabetes based on their health attributes. This involves building a predictive model using historical data with known outcomes (presence or absence of diabetes) to make predictions on new data.

The target variable is the binary indicator of diabetes risk, where individuals are categorized as either high risk (1) or low risk (0) based on the dataset.

Dataset Description

We found two sources of datasets:

- Early stage diabetes risk dataset: This dataset is provided by UCI Repository with comprehensive attributes all in one csv file:
    - diabetes_data_upload.csv: *Age, Sex, Polyuria, Polydipsia, sudden weight loss, weakness, polyphagia, Genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, class*

  https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset
- Diabetes prediction dataset: This dataset is a collection of medical and demographic data from patients containing the following attributes:
    - diabetes_prediction.csv: *gender, age,hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, diabetes*

  https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data

Conclusion

By leveraging data science techniques to analyze health attributes and predict diabetes risk, this project aims to provide valuable insights for insurance companies to effectively assess and manage risks associated with potential customers. Through accurate risk assessment, insurers can optimize their pricing strategies, enhance customer satisfaction, and promote better health outcomes.

The modeling process of this proposal is relatively clear in this proposal. The datasets may have missing value, so simple preparation is needed for the datasets. We are expecting to get high accuracy and dig deeper into the business value of the problem.

Student Performance Prediction

Business analysis

Problem description

As a consulting firm in the educational field, the problem we aim to address is how to optimize student performance and academic success. Specifically, we seek to understand the factors that significantly influence student performance to propose and develop targeted interventions and strategies to improve educational outcomes.

Expected results and actions

Through our analysis, expect to gain insights into the various factors impacting student performance, such as demographic characteristics, socio-economic background, parental involvement, school resources, and teaching methodologies. Additionally, we anticipate uncovering patterns and correlations within the data that can highlight both positive and negative influences on student success.

Based on the results of our analysis, the firm can propose several actionable steps to enhance student performance. These actions may include:

- Implementing targeted intervention programs for students identified as at-risk based on predictive modeling.
- Investing in resources and support systems for schools serving disadvantaged communities.
- Providing professional development opportunities for educators to adopt evidence-based teaching practices.
- Collaborating with policymakers to advocate for educational reforms aimed at addressing systemic barriers to student success.

Data science problem

Prediction target

This is a supervised learning problem, specifically a regression task. We aim to predict student performance (the target variable) based on various attributes such as demographics, socio-economic factors, parental involvement, and school-related variables.

Dataset Description

For this analysis, we propose using the "Student Performance" dataset available from the UCI Machine Learning Repository. This dataset contains comprehensive information about students' background, family, and academic performance, making it suitable for our investigation into the factors influencing student success. The dataset contains 33 attributes and no missing value.

https://archive.ics.uci.edu/dataset/320/student+performance

Conclusion

Through this project, we aim to empower educational stakeholders with actionable insights to drive positive change and improve outcomes for all students. Through a combination of predictive modeling, exploratory analysis, and strategic recommendations, we can work towards creating a more equitable and supportive educational environment.

Twitch Gamer Followers Prediction

Business analysis

Problem description

Nowadays, user-generated contents are taking more and more places than the traditional media. Social network becomes an effective way of promotion targeted at the younger demographic. The number of the followers can serve as an indicator to measure a user's promotional capabilities. However, the promotion through social network need the enterprises develop trust relationships with the streamers. What's more, connecting with the top streamers may cause heavy time and money cost. If the enterprises can find out the potential stars and invest them at early professional stage, it may bring more long-term benefits for both the enterprises and the streamers.

Expected results and actions

In this project we are expecting to evaluate potential followers of the streamers which may support enterprise's commercial decisions. For an new account, we predict whether its future followers' number will grow more than 10000 based on its life circle information and contents. If it does, we determine this account as valuable to invest. The enterprises could set up early connections with those potential valuable accounts with a relatively low cost or support their growth. For the social network agency, they can also use this model to filter out new candidates to make contracts.

Data science problem

Prediction target

This is a supervised learning problem in the field of classification. We aim to classify the accounts as potential and limited. The input is the current accounts' information and the output is a Boolean value.

Dataset Description

We use a single dataset *Twitch Gamers* for this project. The origin dataset is a network. It contains two csv files. One indicates the accounts included in the datasets with labeled attributes, another indicates the edges between them. We will transfer this into a single table by generating a new column of number of followers by calculating the degrees of the nodes.

The dataset contains following attributes: *begin date, life time, latest update time, explicit content, language ,affiliated company*

https://snap.stanford.edu/data/twitch_gamers.html

Conclusion

This project is based on a different type of dataset and it need a little prepare work before the modeling stage. Social network is a trending topic and worth to be dug deeper. We are expecting to find out more business value from the social network data and try to establish innovative business model based on the analysis.