Projects

Real-time Fall Detecting System (Thesis Project)

Tech: Machine Learning, Deep Learning, Transfer Learning, Curriculum Learning, PyTorch;

(1) Object: to predict fall or not for the current frame from web-camera, which is usually used in home-care and healthcare institution where camera installation is permitted;
(2) Adapt transfer learning and curriculum learning into the project;
(3) Introduce data augmentation with threshold searching method to find best threshold to oversample minority class samples ratio;
(4) Meanwhile, datasets splitting, early stop technique, as well as multiple evaluation metrics monitoring the improvement of model, which all of them made mAP increasing 20% even when there were very little open datasets used.
(5) In order to protect the privacy of individuals as well as increasing training data, we not keep the original videos, but keep the new skeleton images;

Medical Insurance Detection based on new National Medical Policy of DRG (Company Project)

Tech: Graph Algorithm, ERNIESage, Paddlepaddle;

(1)Object: to predict whether the medical service items matched the primary and secondary diseases and surgeries for inpatient medical records;
(2)introduce ERNIESage graph algorithm to get the embedding features from all nodes, which not only captures the semantic features of items but also extracts the graph-structural information in which items resides;
(3) do L2 normalization for graph embedding nodes, and then calculate the dot product between items and main disease, secondary disease as well as the surgeries;
(4) Compared to traditional knowledge-graph-based path methods for determining whether a project is related to a disease or procedure, this approach not only overcomes the challenge of approximately 90% of items being unmatched, but also resolves the inconsistency caused by the randomness in path selection that leads to varying model outputs. At the same time, monitoring effectiveness is greatly enhanced, and users from the healthcare insurance bureau can adjust the similarity threshold according to the desired level of strictness.

Monthly Sales Prediction of Alcohol Products on JD Platform (Company Project)

Tech: Machine Learning, Statistics, Python;

(1) Objective: To predict the natural monthly sales of alcoholic products on the JD.com platform.
(2) Challenges: The client is unwilling to provide their actual sales data. Therefore, we inferred sales using the number of reviews on JD.com. The corresponding mathematical derivations and parameter extraction methods are detailed on GitHub.
(3) First-layer framework: Beyond field selection, data cleaning, reconstruction, and imputation, we primarily employed a sliding-window approach to construct features reflecting the recent changes in review counts for each product and its associated group, along with historical statistics. These features were then used to train a LightGBM model with 5-fold cross-validation.
(4) Second-layer framework: Predicted review counts were converted into natural monthly sales. Since there is no direct sales data for alcoholic products, we applied business logic and established mathematical formulas to map the relationship between review counts and sales observed on similar Platform Taobao to the JD.com platform.
(5) Outcome: Although the client could not provide actual sales data, this approach served as an exploratory method. It was ultimately adopted, deployed on the platform, and integrated into monthly reports.

Entity Identification of Chinese and English Brand Names for Cosmetics Products (Company Project)

Tech: Machine Learning, Deep Learning, NLP, NER, PyTorch;

(1) Objective: To predict the Chinese and English brand names of cosmetics mentioned in the samples.

(2) Data preprocessing: Approximately 70% of the samples were covered using the company's existing keyword-matching rules. Based on business requirements, samples where a brand appeared three or more times were selected as training data, while the remaining 30% unmatched samples were used as the test set.
(3) Annotation strategy: Instead of relying on third-party manual annotation, an automated Python-based NER labeling pipeline was developed, significantly saving manpower, financial resources, and time.
(4) Model framework: BiLSTM+CRF was introduced, with both BIO+40pad and BIEO+40pad labeling schemes were tested. It was observed that the two labeling approaches provided complementary predictions, with BIEO achieving slightly higher precision.
(5) For model evaluation, business-driven accuracy was used: if at least one brand in a sample was correctly predicted, the prediction was considered correct. The model achieved an accuracy of approximately 85%.

Regression Prediction of the Next Maintenance Date and Kilometers (Company Project)

Tech: Machine Learning, Statistics, Python.

(1) Based on business data analysis, all users were divided into two groups: 70% were predicted using machine learning models, while 30% were predicted using statistical methods.
(2) For the model-based predictions, the feature engineering focused on extracting single-dimensional and cross-dimensional statistical feature sets of users and their corresponding groups, including maintenance days, mileage, and daily averages.
(3) An innovative training approach was adopted: to predict the (n+1) date and mileage for those customers who has historical data of length n by training those who have historical data of length (n+1).
(4) For statistical predictions, the user groups were finely divided into eight categories according to data characteristics, enabling more granular forecasts.
(5) Results: Over 40% of users had a predicted time deviation within one month using the machine learning model, while statistical predictions achieved 31%, only about 10% lower. Based on these outcomes, business or marketing teams can implement targeted promotions or distribute coupons to drive sales, while the same data can also be leveraged to manage customer attrition effectively.

Risks Prediction for Applications Changes in Citi Bank of China (Company Project)

Tech: Machine Learning, Statistics, Python.

(1) Objective: To predict the risk associated with all change implementations within the GCG and ICG departments.
(2) In addition to GCG department data, ICG department data was incorporated to address the issue of limited sample size. At the same time, the original simple linear regression model was replaced with a binary classification model to predict the risk of app changes. Furthermore, basic app attributes were extracted from the bank's open system as supplementary features.
(3) Feature engineering: Constructed statistical feature groups by aggregating counts and failure frequencies across temporal dimensions, based on both single-dimensional and cross-dimensional combinations of factors such as applicator, change content, application, change window, app environment, app vendor, and app user department.
(4) Feature selection & modeling: Testing showed that retaining all features produced optimal results, so no filtering was applied. The final model adopted a weighted integration of LightGBM (LGB) and Logistic Regression (LR).
(5) Practical deployment: For changes identified as high risk, additional segmentation by risk interval was applied, with differentiated approval workflows and approval levels defined for each segment. This approach effectively reduced the overall number of changes by approximately 20%, while extremely high-risk changes were jointly reviewed within the International Change Management process, thereby mitigating the likelihood of change failures.
(6) At this point, it becomes evident that when machine learning algorithms are integrated with business processes, the resulting value can be immense.