Projects

Real-time Fall Detecting System (Thesis Project)

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

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

(1) Object: to predict whether the medical service items matched the primary and secondary diseases and surgeries from 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 reside.
(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 in JD Platform (Company Project)

(1) Objective: To use dynamic comment number data from the JD platform to predict and infer the monthly sales of all alcoholic items on the JD platform
(2) The primary difficulty: predicting sales using comment data—it is uncertain whether the results are usable or meaningful. Also, the comment dataset is relatively small and contains a high prevalence of noisy data etc.
(3) First-layer framework: After data preprocessing and statistical analysis, we segmented our selected customer data into three groups, each primarily employed a sliding-window approach with different strides to construct features engineering and then train and predict by LightGBM model with different groups.
(4) Second-layer framework: Predicted comment counts were converted into natural monthly sales. Per there is no direct sales data for alcoholic products, we applied business logic and established mathematical formulas to map the relationship between comment numbers and sales.
(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.
(6) Welcome to visit my GitHub to explore the project task breakdown, in-depth analyses, and the detailed design of algorithmic model frameworks.

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

(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 being 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 all Customers' Next Car-service Date and Odometer (Company Project)
(1) Transformed the original long-format dataset into a wide-format structure to prevent data leakage during training, avoiding inflated model performance.
(2) Fed 70% of well-cleaned data into the algorithmic model, while the remaining 30%—including cold-start users or data that could not be fully cleaned—were processed using a statistical rule-based model.
(3) Both algorithmic and statistical rule-based models innovatively employed granular and batch-wise training and statistical methods.
(4) Within the Machine Learning model, an innovative multi-metric weighted fusion approach was applied, enabling strong performance even with a single model.
(5) Results: About 41% 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, our VIP client requested to deploy with PySpark in production in their company and important sub-companies.
(6) Welcome to visit my GitHub to explore the project task breakdown, in-depth analyses, and the detailed design of algorithmic model frameworks.

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