**Home Credit包装简历**

1 本次竞赛由Home Credit Group发起的，该公司是一家国际非银行金融机构。这个竞赛的目标是预测哪些客户更有可能违约贷款。

2 基于比赛数据进行特征工程。比赛方提供个人基本资料（性别、年龄、学历等），消费记录（每月消费金额、限制消费原因等），信用卡资料（开户行、级别等）等465个原生的特征。我们基于这些特征考虑如下衍生特征：

* 聚合特征：统计每个用户消费水平的聚合特征，分别从某个角度反映该用户的情况。如果一个变量是连续型，我们可以取mean、std、min、max；如果一个变量是离散型，我们可以取count、last和nunique。
* Diff特征：统计用户的历史消费数据的变化水平，该特征可以反映用户近期是否产生了行为模式上的变化。

3 在风险控制问题中，决策树模型通常有较好的表现，这里我们采用LightGBM和CatBoost模型来解决这个问题。该类模型都已经在多种任务中显示出了卓越的性能，也很契合本次比赛的数据特点。为了提升模型的稳健性，我们参考样本的week\_num使用groupkfold划分成五段测试集充分测试，避免过拟合。

4 在比赛中，我们充分利用比赛的时间，合理控制比赛节奏；积极进行讨论，制定了合理的比赛模型迭代任务，并组织了头脑风暴来推进比赛进展。同时，大家着手进行了模型的开发工作，确保能够尽可能地提高预测准确性。最终我们得分为xxx，取得了top x%的排名。

1. This competition was initiated by Home Credit Group, an international non-banking financial institution. The goal of this competition is to predict which customers are more likely to default on loans.

2. Feature engineering was conducted based on the competition data. The competition provided 465 native features including personal information (gender, age, education, etc.), consumption records (monthly consumption amount, reasons for consumption restrictions, etc.), and credit card information (issuing bank, level, etc.). We derived the following features based on these characteristics:

- Aggregate Features: Aggregate features that summarize the consumption level of each user, reflecting the user's situation from different perspectives. For a continuous variable, we can take mean, std, min, max; for a discrete variable, we can take count, last, and nunique.

- Diff Features: Statistic of the change in the user's historical consumption data, which can reflect whether there has been a change in behavioral patterns in the recent period.

3. In risk control issues, decision tree models usually perform well. Here, we used LightGBM and CatBoost models to solve this problem. These types of models have demonstrated excellent performance in various tasks and are well suited to the data characteristics of this competition. To improve the robustness of the model, we used groupkfold to divide the samples into five segments of test sets based on the week\_num of the samples for thorough testing and to avoid overfitting.

4. In the competition, we made full use of the competition time, controlled the pace of the competition reasonably, actively discussed and formulated reasonable competition model iteration tasks, and organized brainstorming sessions to promote the progress of the competition. At the same time, we started developing the models to ensure that we could improve predictive accuracy as much as possible. In the end, we scored xxx points, achieving a ranking in the top x%.