Predicting churn using machine learning, and more specifically deep learning models, hinges on the quality of the input data. Things like age, location, and subscription info help us group users and personalize their experience—but if that data’s inconsistent or messy, the model can’t really make sense of it or spot clear patterns.

**Data Quality Issues in Customer Profiles**

1. **Missing Age Values**  
   Customer base lacks age information. This restricts the model’s ability to identify churn patterns by age group.
2. **Inconsistent Location Formats**  
   Geographic entries vary (e.g., “TX,” “Texas,” “Tex.”), leading to data being grouped incorrectly by region, which makes it harder to spot trends based on where users live.
3. **Subscription Type Variability**  
   Multiple representations of the same subscription level appear in the dataset, such as “Free,” “free-tier,” or “Basic-Free,” introducing unnecessary category inflation.
4. **Null and Invalid Payment Methods**  
   Some entries contain unrecognized or blank payment types, which confuses transaction-related churn analysis.
5. **Outdated Subscription Status**  
   In some profiles, the subscription status hasn’t been updated to reflect recent payment behavior, causing inaccuracies in current churn labels.

**Preprocessing Techniques and Justifications**

1. **Age Imputation Using Regional Medians**  
   Missing ages will be filled using the median age of customers within the same location. Ensure Date of Birth entry is mandatory moving forward in order to improve churn analysis.
2. **Location Standardization via Mapping Dictionary**  
   Tools like the CDX Technologies ZIP+4 Lookup or USPS’s Address Element Correction (AEC) service can take in partial or messy address data and return a clean, standardized version (TX).
3. **Subscription Label Normalization**  
   All variations of a subscription type will be standardized (e.g., “free-tier,” “basic-free” → “Free”) to make the data representations more compact and help the model perform better.
4. **Payment Method Encoding with “Unknown” Category + App notifications**  
   Valid payment types will be one-hot encoded. Null/invalid entries will be grouped into a separate “Unknown” bucket to prevent data loss while preserving representation. The app notifications would allow the member to be aware there is a payment issue that needs to be resolved.
5. **Cross-referencing with Billing Logs for Accuracy**  
   Subscription status will be validated against recent transaction data and corrected where necessary. This ensures that training labels represent real-world behaviors.

**Conclusion**

Structured customer profile data can be a powerful predictor of churn—if properly cleaned. The data issues identified here affect a deep learning model's accuracy, bias, and interpretability.

**References**

Jadhav, A., Kumar, V., & Joshi, A. (2019). A review on missing data imputation using machine learning techniques. *Procedia Computer Science, 160*, 803–810. <https://doi.org/10.1016/j.procs.2019.11.027>

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin, 23*(4), 3–13